Universality for generalized Wigner matrices with Bernoulli distribution

László Erdős[1]*, Horng-Tzer Yau[2]† and Jun Yin[2]

Institute of Mathematics, University of Munich,
Theresienstr. 39, D-80333 Munich, Germany
lerdos@math.lmu.de [1]

Department of Mathematics, Harvard University
Cambridge MA 02138, USA
htyau@math.harvard.edu, jyin@math.harvard.edu [2]

Aug 11, 2010

Abstract

The universality for the eigenvalue spacing statistics of generalized Wigner matrices was established in our previous work [19] under certain conditions on the probability distributions of the matrix elements. A major class of probability measures excluded in [19] are the Bernoulli measures. In this paper, we extend the universality result of [19] to include the Bernoulli measures so that the only restrictions on the probability distributions of the matrix elements are the subexponential decay and the normalization condition that the variances in each row sum up to one. The new ingredient is a strong local semicircle law which improves the error estimate on the Stieltjes transform of the empirical measure of the eigenvalues from the order $(N\eta)^{-1/2}$ to $(N\eta)^{-1}$. Here η is the imaginary part of the spectral parameter in the definition of the Stieltjes transform and N is the size of the matrix.

AMS Subject Classification: 15A52, 82B44

Keywords: Random band matrix, Local semicircle law, sine kernel.

*Partially supported by SFB-TR 12 Grant of the German Research Council
†Partially supported by NSF grants DMS-0757425, 0804279

1 Introduction

The universality of local eigenvalue statistics in the bulk of the spectrum of random matrices has been traditionally considered only for invariant ensembles [4, 7, 8, 25]. For non-invariant ensembles, a new approach to prove the bulk universality was developed in [16, 14, 18, 19]. It consists of the following three steps:

1. Local semicircle law.
2. Universality for Gaussian divisible ensembles.
3. Approximation by Gaussian divisible ensembles.

In Step 2, the universality of the local eigenvalue statistics for a large class of matrices, i.e., Gaussian divisible matrices, was established. Thus in order to prove the universality of a given ensemble, it remains to approximate the matrix elements in this ensemble by Gaussian divisible distribution in such a way that the local eigenvalue statistics are unchanged. This approximation is intrinsically a density theorem and it can be achieved by perturbative expansions in several different ways. In the most recent approach [18, 19], the universality for Gaussian divisible ensembles was proved via the Dyson Brownian motion and the stability of eigenvalues in Step 3 was provided by the Green function comparison theorem. In Step 2 a technical tool, the logarithmic Sobolev inequality (LSI), was needed to estimate the fluctuations of eigenvalue distribution. This restriction could not be completely removed in Step 3 and thus the Bernoulli measures were excluded in [19]. In this paper, we will improve the local semicircle law so that the LSI is no longer needed. This will enable us to prove the universality for generalized Wigner matrices with Bernoulli distributions. As a byproduct of the new stronger form of local semicircle law, we also obtain much stronger estimates on the eigenvalue density and on the matrix elements of the resolvent.

Recall the Stieltjes transform of the empirical measure of the eigenvalues $\{\lambda_j\}_{j=1}^N$ is defined by

$$m_N(z) = \frac{1}{N}\sum_{j=1}^N \frac{1}{\lambda_j - z}.$$

We have proved in [19] that the difference between $m_N(z)$ and $m_{sc}(z)$, the Stieltjes transform of the semicircle law (2.9), is bounded by $(N\eta)^{-1/2}$ where $\eta = \mathfrak{Im}\, z$. The main result of this paper states that the error can be improved to $(N\eta)^{-1}$. The improvement of a factor $(N\eta)^{-1/2}$ resembles the usual $N^{-1/2}$ factor in the central limit theorem and it results from a new estimate on the correlations of error terms. This estimate also implies that the error between the normalized empirical counting function of the eigenvalues and the one given by the semicircle law is less than $N^{-1+\varepsilon}$ in the bulk of the spectrum for any $\varepsilon > 0$. This new input is sufficiently strong to replace the usage of the (LSI) in [19], see the discussion after Theorem 2.2 for more details.

Notice that this improvement of a factor $(N\eta)^{-1/2}$ and the removal of the LSI need a substantial amount of work. Our motivations to take on this endeavor are for the following two reasons: (1) The distributions of the Bernoulli random matrices are very singular while the Gaussian measures in GOE are very smooth. It is not a priori clear that the universality holds for such singular distributions. (2) The adjacency matrices for random graphs are natural examples of symmetric random matrices. The matrix elements of these matrices take the values 0 or 1 and thus they form Bernoulli random matrices. Our current results do not cover this case since we require the mean zero condition, but they represent the first step toward the universality of the adjacency matrices of random graphs.

2 Main results

We now state the main results of this paper. Since all our results hold for both hermitian and symmetric ensembles, we will state the results for the hermitian case only. The modifications to the symmetric case are straightforward and they will be omitted. Let $H = (h_{ij})_{i,j=1}^N$ be an $N \times N$ hermitian matrix where the matrix elements $h_{ij} = \overline{h}_{ji}$, $i \le j$, are independent random variables given by a probability measure ν_{ij} with mean zero and variance σ_{ij}^2. The variance of h_{ij} for $i > j$ is $\sigma_{ij}^2 = \mathbb{E}\,|h_{ij}|^2 = \sigma_{ji}^2$. For simplicity of the presentation, we assume that for any fixed $1 \le i < j \le N$, $\mathrm{Re}\, h_{ij}$ and $\mathrm{Im}\, h_{ij}$ are i.i.d. with distribution ω_{ij}, i.e., $\nu_{ij} = \omega_{ij} \otimes \omega_{ij}$ in the sense that $\nu_{ij}(\mathrm{d}h) = \omega_{ij}(\mathrm{dRe}\, h)\omega_{ij}(\mathrm{dIm}\, h)$, but this assumption is not essential for the result. The distribution ν_{ij} and its variance σ_{ij}^2 may depend on N, but we omit this fact in the notation. We assume that for any j fixed

$$\sum_i \sigma_{ij}^2 = 1\,. \tag{2.1}$$

Matrices with independent, zero mean entries and with the normalization condition (2.1) will be called *universal Wigner matrices.* The basic parameter of such matrices is the quantity

$$M := \frac{1}{\max_{ij} \sigma_{ij}^2}. \tag{2.2}$$

Define C_{inf} and C_{sup} by

$$C_{inf} := \inf_{N,i,j} \{N\sigma_{ij}^2\} \le \sup_{N,i,j} \{N\sigma_{ij}^2\} =: C_{sup}. \tag{2.3}$$

Note that $C_{inf} = C_{sup} (= 1)$ corresponds to the standard Wigner matrices and the conditions $0 < C_{inf} \le C_{sup} < \infty$ define more general Wigner matrices with comparable variances.

We will also consider an even more general case when σ_{ij} for different (i,j) indices are not comparable. A special case is the *band matrix*, where $\sigma_{ij} = 0$ for $|i - j| > W$ with some parameter W.

Denote by $\Sigma := \{\sigma_{ij}^2\}_{i,j=1}^N$ the matrix of variances which is symmetric, doubly stochastic by (2.1), and in particular satisfies $-1 \le \Sigma \le 1$. Let the spectrum of Σ be supported in

$$\mathrm{Spec}(\Sigma) \subset [-1 + \delta_-, 1 - \delta_+] \cup \{1\} \tag{2.4}$$

with some nonnegative constants $\delta_\pm$. We will always have the following spectral assumption

$$\textit{1 is a simple eigenvalue of } \Sigma \textit{ and } \delta_- \textit{ is a positive constant, independent of } N. \tag{2.5}$$

The local semicircle law will be proven under this general condition, but the precision of the estimate near the spectral edge will also depend on δ_+ in an explicit way. For the orientation of the reader, we mention two special cases that provided the main motivation for our work.

One important class of universal Wigner matrices is the *generalized Wigner ensemble* which is defined by the extra condition that

$$0 < C_{inf} \le C_{sup} < \infty, \tag{2.6}$$

It is easy to check that (2.4) holds with

$$\delta_\pm \ge C_{inf}. \tag{2.7}$$

Another example is the *band matrix ensemble* whose variances are given by

$$\sigma_{ij}^2 = W^{-1} f\Big(\frac{[i-j]_N}{W}\Big), \tag{2.8}$$

where $W \geq 1$, $f : \mathbb{R} \to \mathbb{R}_+$ is a nonnegative symmetric function with $\int f = 1$, $f \in L^\infty(\mathbb{R})$, and we defined $[i-j]_N \in \{1, 2, \ldots N\}$ by the property that $[i-j]_N \equiv i - j \bmod N$. The bandwidth M defined in (2.2) satisfies $M \leq W/\|f\|_\infty$. In Appendix A of [19], we have proved that (2.5) is satisfied for the choice of (2.8) if W is large enough.

Define the Stieltjes transform of the empirical eigenvalue distribution of H by

$$m(z) = m_N(z) := \frac{1}{N} \text{Tr} \, \frac{1}{H-z}, \quad z = E + i\eta.$$

Define $m_{sc}(z)$ as the unique solution of

$$m_{sc}(z) + \frac{1}{z + m_{sc}(z)} = 0,$$

with positive imaginary part for all z with Im $z > 0$, i.e.,

$$m_{sc}(z) = \frac{-z + \sqrt{z^2 - 4}}{2}. \tag{2.9}$$

Here the square root function is chosen with a branch cut in the segment $[-2, 2]$ so that asymptotically $\sqrt{z^2-4} \sim z$ at infinity. This guarantees that the imaginary part of m_{sc} is non negative for Im $z > 0$ and it is the Wigner semicircle distribution

$$\varrho_{sc}(E) := \lim_{\eta \to 0+0} \frac{1}{\pi} \mathfrak{Im}\; m_{sc}(E + i\eta) = \frac{1}{2\pi}\sqrt{(4 - E^2)_+}. \tag{2.10}$$

The Wigner semicircle law [32] states that $m_N(z) \to m_{sc}(z)$ for any fixed z, i.e., provided that η is independent of N. We have proved [19] a local version of this result for universal Wigner matrices and the main result can be stated as the following probability estimate:

$$\mathbb{P}\left(|m_N(z) - m_{sc}(z)| \geq (\log N)^{C_2} \frac{1}{\sqrt{M\eta}\,\kappa} \right) \leq C N^{-c(\log\log N)}$$

with some constant C_2. The accuracy of this estimate can be improved from $(M\eta)^{-1/2}\,\kappa^{-1}$ to $(M\eta)^{-1}\,\kappa^{-1}$, which is the content of the next theorem. It summarizes the results of Theorems 4.1 and 5.1. Prior to our result in [19], a central limit theorem for the semicircle law on macroscopic scale for band matrices was established by Guionnet [21] and Anderson and Zeitouni [2]; a semicircle law for Gaussian band matrices was proved by Disertori, Pinson and Spencer [9]. For a review on band matrices, see the recent article [27] by Spencer.

Theorem 2.1 (Local semicircle law) *Let H be a hermitian $N \times N$ random matrix with $\mathbb{E}\, h_{ij} = 0$, $1 \leq i, j \leq N$, and assume that the variances σ_{ij}^2 satisfy* (2.1) *and* (2.5)*. Suppose that the distributions of the* *matrix elements have a uniformly subexponential decay in the sense that there exist constants α, $\beta > 0$, independent of N, such that for any $x > 0$ we have*

$$\mathbb{P}(|h_{ij}| \geq x^{\alpha}|\sigma_{ij}|) \leq \beta e^{-x}. \tag{2.11}$$

We consider universal Wigner matrices and its special class, the generalized Wigner matrices in parallel. The parameter A will distinguish between the two cases; we set $A = 2$ for universal Wigner matrices, and $A = 1$ for generalized Wigner matrices, where the results will be stronger.

Define the following domain in $\mathbb{C}$

$$D := \Big\{ z = E + i\eta \in \mathbb{C} \; : \; |E| \leq 5, \; 0 < \eta \leq 10, \;\; \sqrt{M\eta} \geq (\log N)^{C_1}(\kappa+\eta)^{\frac{1}{4}-A} \Big\} \tag{2.12}$$

where $\kappa := \big||E| - 2\big|$. Then there exist constants C_1, C_2, C and $c > 0$, depending only on α, β and δ_- in (2.5)*, such that for any $\varepsilon > 0$ and $K > 0$ the Stieltjes transform of the empirical eigenvalue distribution of H satisfies*

$$\mathbb{P}\left(\bigcup_{z \in D} \Big\{ |m_N(z) - m_{sc}(z)| \geq \frac{N^{\varepsilon}}{M\eta\,(\kappa+\eta)^{A}} \Big\} \right) \leq \frac{C(\varepsilon, K)}{N^{K}} \tag{2.13}$$

for sufficiently large N. Furthermore, the diagonal matrix elements of the Green function $G_{ii}(z) = (H - z)^{-1}(i,i)$ satisfy that

$$\mathbb{P}\left(\bigcup_{z \in D} \Big\{ \max_i |G_{ii}(z) - m_{sc}(z)| \geq \frac{(\log N)^{C_2}}{\sqrt{M\eta}}\, (\kappa+\eta)^{\frac{1}{4}-\frac{A}{2}} \Big\} \right) \leq CN^{-c(\log\log N)} \tag{2.14}$$

and for the off-diagonal elements we have

$$\mathbb{P}\left(\bigcup_{z \in D} \Big\{ \max_{i \neq j} |G_{ij}(z)| \geq \frac{(\log N)^{C_2}}{\sqrt{M\eta}}\, (\kappa+\eta)^{\frac{1}{4}} \Big\} \right) \leq CN^{-c(\log\log N)} \tag{2.15}$$

for any sufficiently large N.

The subexponential decay condition (2.11) can also be easily weakened if we are not aiming at error estimates faster than any power law of N. This can be easily carried out and we will not pursue it in this paper.

Denote the eigenvalues of H by $\lambda_1, \ldots, \lambda_N$ and let $p_N(\lambda_1, \ldots, \lambda_N)$ be their (symmetric) probability density. For any $k = 1, 2, \ldots, N$ the k-point correlation function of the eigenvalues is defined by

$$p_N^{(k)}(x_1, x_2, \ldots x_k) := \int_{\mathbb{R}^{N-k}} p_N(x_1, x_2, \ldots, x_N) \mathrm{d}x_{k+1} \ldots \mathrm{d}x_N. \tag{2.16}$$

We now state our main result concerning these correlation functions. The same result was proved in [19] under the additional assumption (2.26).

Theorem 2.2 (Universality for generalized Wigner matrices) *Consider a generalized hermitian Wigner ensemble such that* (2.1)*,* (2.5) *and* (2.6) *hold. Suppose that the distributions ν_{ij} of the matrix elements have a uniformly subexponential decay in the sense of* (2.11)*. Suppose that the real and imaginary* *parts of h_{ij} are i.i.d., distributed according to ω_{ij}, i.e., $\nu_{ij}(\mathrm{d}h) = \omega_{ij}(\mathrm{d}\mathfrak{Im}\, h)\omega_{ij}(\mathrm{d}\mathfrak{Re}\, h)$. Then for any $k \geq 1$ and for any compactly supported continuous test function $O : \mathbb{R}^k \to \mathbb{R}$ we have*

$$\begin{aligned} \lim_{b\to 0}\lim_{N\to\infty} \frac{1}{2b}\int_{E-b}^{E+b} \mathrm{d}E' \int_{\mathbb{R}^k} & \mathrm{d}\alpha_1 \dots \mathrm{d}\alpha_k \; O(\alpha_1, \dots, \alpha_k) \\ & \times \frac{1}{\varrho_{sc}(E)^k}\Big(p_N^{(k)} - p_{GUE,N}^{(k)}\Big)\Big(E' + \frac{\alpha_1}{N\varrho_{sc}(E)}, \dots, E' + \frac{\alpha_k}{N\varrho_{sc}(E)}\Big) = 0, \end{aligned} \tag{2.17}$$

where $p_{GUE,N}^{(k)}$ is the k-point correlation function for the GUE ensemble. The same statement holds for symmetric matrices, with GOE replacing the GUE ensemble.

Remark. We can take $b = N^{-c}$ for some small constant $c > 0$ so that there is no double limit taken. This is because all our bounds have an effective error estimate N^{-c}. In case of hermitian matrices there is no need for averaging in the energy parameter E'. The limit (2.17) holds even for any fixed energy E', with $|E'| < 2$, since, instead of relying on the local relaxation flow of [14, 18], we can use the result of [16] for Gaussian divisible ensembles at a fixed energy.

It is well-known that the limiting correlation functions of the GUE ensemble are given by the sine kernel

$$\frac{1}{\varrho_{sc}(E)^k} p_{GUE,N}^{(k)}\Big(E + \frac{\alpha_1}{N\varrho_{sc}(E)}, \dots, E + \frac{\alpha_k}{N\varrho_{sc}(E)}\Big) \to \det\{K(\alpha_i - \alpha_j)\}_{i,j=1}^k, \qquad K(x) = \frac{\sin \pi x}{\pi x},$$

and a similar universal formula is available for the limiting gap distribution. The formulas for the GOE cases are more complicated and we refer the reader to standard references such as [1, 6, 20, 24].

We will prove Theorem 2.2 using the approach of [18, 19]. The logarithmic Sobolev inequality was an important tool in these papers and it was the main obstacle why the case of Bernoulli random matrices were not covered. We note that the Bernoulli distribution satisfies the discrete version of the LSI but it would not be sufficient for our purposes. To explain the necessity of LSI, we now review the three basic ingredients of the approach of [18, 19].

Step 1. *Local semicircle law:* It states that the density of eigenvalues is given by the semicircle law down to short scales containing only N^ε eigenvalues for all $\varepsilon > 0$, where N is the size of the matrix.

Step 2. *Local ergodicity of the Dyson Brownian motion:* The Dyson Brownian motion is given by the flow

$$H_t = e^{-t/2} H_0 + (1 - e^{-t})^{1/2}\, V, \tag{2.18}$$

where H_0 is the initial Wigner matrix, V is an independent standard GUE (or GOE) matrix and $t \geq 0$ is the time. Here we have used the version that the dynamics of the matrix element is given by an Ornstein-Uhlenbeck (OU) process on $\mathbb{C}$. More precisely, let

$$\mu = \mu_N(\mathrm{d}\mathbf{x}) := \frac{e^{-\mathcal{H}(\mathbf{x})}}{Z_\beta}\mathrm{d}\mathbf{x}, \qquad \mathcal{H}(\mathbf{x}) = \mathcal{H}_N(\mathbf{x}) := N\left[\beta\sum_{i=1}^N \frac{x_i^2}{4} - \frac{\beta}{N}\sum_{i<j}\log|x_j - x_i|\right] \tag{2.19}$$

be the probability measure of the eigenvalues $\mathbf{x} = (x_1, x_2, \dots, x_N)$ of the general β ensemble, $\beta \geq 1$ ($\beta = 2$ for the hermitian case and $\beta = 1$ for the symmetric case). Denote the distribution of the eigenvalues of H_t at time t by $f_t(\mathbf{x})\mu(\mathrm{d}\mathbf{x})$. Then $f_t = f_{t,N}$ satisfies [10]

$$\partial_t f_t = \mathscr{L} f_t. \tag{2.20}$$

where

$$\mathscr{L} = \mathscr{L}_N := \sum_{i=1}^{N} \frac{1}{2N}\partial_i^2 + \sum_{i=1}^{N} \bigg(-\frac{\beta}{4}x_i + \frac{\beta}{2N}\sum_{j\neq i}\frac{1}{x_i - x_j}\bigg)\partial_i. \tag{2.21}$$

We now recall the following theorem concerning the universality of the Dyson Brownian motion. Following the convention in [18], we label the assumptions as Assumptions II–IV since the Assumption I, a convexity property of the Hamiltonian for the invariant measure of the Dyson Brownian motions, is automatically satisfied for any β ensembles.

Assumption II. For any fixed $a, b \in \mathbb{R}$, we have

$$\lim_{N\to\infty}\sup_{t\geq 0}\bigg| \int \frac{1}{N}\sum_{j=1}^{N} \mathbf{1}(x_j \in [a,b]) f_t(\mathbf{x})\mathrm{d}\mu(\mathbf{x}) - \int_a^b \varrho_{sc}(x)\mathrm{d}x \bigg| = 0. \tag{2.22}$$

where ϱ_{sc} is the density of the semicircle law (2.10).

Let $\gamma_j = \gamma_{j,N}$ denote the location of the j-th point under the semicircle law, i.e., γ_j is defined by

$$N\int_{-\infty}^{\gamma_j} \varrho_{sc}(x)\mathrm{d}x = j, \qquad 1 \leq j \leq N. \tag{2.23}$$

We will call γ_j the *classical location* of the j-th point.

Assumption III. There exists an $\varepsilon > 0$ such that

$$\sup_{t\geq 0}\int \frac{1}{N}\sum_{j=1}^{N}(x_j - \gamma_j)^2 f_t(\mathrm{d}\mathbf{x})\mu(\mathrm{d}\mathbf{x}) \leq CN^{-1-2\varepsilon} \tag{2.24}$$

with a constant C uniformly in N.

The final assumption is an upper bound on the local density. For any $I \in \mathbb{R}$, let

$$\mathcal{N}_I := \sum_{i=1}^{N} \mathbf{1}(x_i \in I)$$

denote the number of eigenvalues in I.

Assumption IV. For any compact subinterval $I_0 \subset (-2,2) = \{E \,:\, \varrho_{sc}(E) > 0\}$, and for any $\delta > 0$, $\sigma > 0$ there are constant C_n, $n \in \mathbb{N}$, depending on I_0, δ and σ such that for any interval $I \subset I_0$ with $|I| \geq N^{-1+\sigma}$ and for any $K \geq 1$, we have

$$\sup_{\tau \geq N^{-2\varepsilon+\delta}} \int \mathbf{1}\big\{\mathcal{N}_I \geq KN|I|\big\} f_\tau \mathrm{d}\mu \leq C_n K^{-n}, \qquad n = 1, 2, \ldots, \tag{2.25}$$

where ε is the exponent from Assumption III and σ and δ are arbitrarily small numbers.

We have proved [19] that Assumption IV follows from the local semicircle law and Assumption III also follows from the local semicircle law provided that a uniform LSI for the distributions of the matrix elements is assumed.

Step 3. *Green function comparison theorem:* It asserts that the correlation functions of the eigenvalues of two matrix ensembles are identical up to the scale $1/N$ provided that the first four moments of the matrix elements of these two ensembles are almost identical. Given this theorem and the universality for the Dyson Brownian motion for $t \sim N^{-\varepsilon}$, the universality for a matrix ensemble H holds if we can find another matrix ensemble H_0 such that the first four moments of the matrix elements of H and H_t (given by (2.18)) are almost the same. Furthermore, H_0 is required to satisfy a uniform LSI so that the Assumption III can be verified. This is possible if the first four moments of H_0 satisfy

$$\inf_N \min_{1\leq i,j\leq N} \left\{ \frac{m_4(i,j)}{(m_2(i,j))^2} - \frac{(m_3(i,j))^2}{(m_2(i,j))^3} \right\} > 1, \tag{2.26}$$

where $m_k(i,j)$ is the k-th moment of the i,j matrix element in the symmetric case. In the hermitian case, the moments of the real and imaginary parts have to satisfy (2.26).

Combining these ingredients, the universality of local eigenvalue statistics in the bulk was proved for all generalized Wigner ensembles (see (2.6) for the definition) satisfying (2.26) and a subexponential decay technical condition. The restriction (2.26) was needed to guarantee the existence of a matching matrix ensemble whose matrix element distributions satisfy the LSI so that the Assumption III can be verified. The local semicircle estimates in Theorem 2.1 imply that the empirical counting function of the eigenvalues is close to the semicircle counting function (Theorem 6.3) and that the location of the eigenvalues are close to their classical location in mean square deviation sense (Theorem 7.1). This provides a direct proof to the Assumption III (2.24) and thus removes the usage of the LSI.

Finally we summarize the recent results related to the bulk universality of local eigenvalue statistics. The local semicircle law for Step 1 was first established for Wigner matrices in a series of papers [11, 12, 13]. The method was based on a self-consistent equation for the Stieltjes transform of the eigenvalues and the continuity of the imaginary part of the spectral parameter in the Stieltjes transform. As a by-product, an eigenvector delocalization estimate was proved.

The universality for Gaussian divisible ensembles was proved by Johansson [23] for *hermitian* Wigner ensembles. It was extended to *complex* sample covariance matrices by Ben Arous and Péché [3]. There were two major restrictions of this method: 1. The Gaussian component was fairly large, it was required to be of order one independent of N. 2. It relies on explicit formulas for the correlation functions of eigenvalues which are valid only for Gaussian divisible ensembles with unitary invariant Gaussian component. The size of the Gaussian component was reduced to $N^{-1+\varepsilon}$ in [16] by using an improved formula for correlation functions and the local semicircle law from [11, 12, 13]. The Gaussian component was then removed by a perturbation argument using the reverse heat flow. Thus the three step strategy to prove the universality was introduced and it led to the first proof of the bulk universality for hermitian Wigner ensembles. Due to the reverse heat flow argument used in Step 3, the universality class established in [16] was restricted to matrices with smooth distributions for the matrix elements. Shortly after, Tao and Vu [28] proved the four moment theorem which in particular removes the smoothness restriction in Step 3. It thus proved the universality for hermitian Wigner matrices whose matrix element distributions were supported on at least three points. The last condition was removed in [17] by combining the arguments of [16, 28]. The result of [28] also implies that the local statistics of symmetric Wigner matrices and GOE are the same, but under the restriction that the first four moments of the matrix elements match those of GOE. Thus the universality class for the local correlation functions established via the approach of combining [28] and [23] was broader for the hermitian ensembles than for the symmetric ones. This improvement was due to Johansson's result [23], which provided the universality for Gaussian divisible ensembles in Step 2, was available only for hermitian ensembles.

A more general and conceptually very appealing approach for Step 2 is via the local ergodicity of Dyson Brownian motion. This approach, initiated in [14], was applied to prove the universality for symmetric Wigner matrices with the three point support condition. In [18], we formulated a general theorem for the bulk universality which applies to all classical ensembles, i.e., real and complex Wigner matrices, real and complex sample covariance matrices and quaternion Wigner matrices. Later on, Tao and Vu [29] also extended their results to the sample covariance matrices with the three point support condition for complex covariance matrices and four moment matching conditions for real ones. Shortly after [29], Péché [26] also extended the approach [16] to the complex sample covariance matrices and proved the universality in the bulk.

Most recently, we introduced [19] the Green function comparison theorem and extended the local semicircle law to include the matrix elements of the Green functions. This allows us to remove the smoothness restriction from the reverse heat flow argument in Step 3 of our approach. We remark that the comparison theorems in [28] concern individual eigenvalues with a fixed index, while the Green function comparison theorem is at a fixed energy. On the other hand, in [19] the variances of the matrix elements were allowed to vary, i.e., the matrices belonged to generalized Wigner ensembles. The three step strategy can thus be applied and the universality was proved for generalized Wigner ensembles with essentially only one class of measures, the Bernoulli measures, excluded due to the LSI used in verifying Assumption III in Step 2. Finally, in the current paper, Assumption III will be shown to be a consequence of a strong local semicircle law, which will be proved for all ensembles with a subexponential decay property. In particular, Bernoulli measures are now included in the universality class (in the sense of (2.17)) for both hermitian and symmetric generalized Wigner ensembles. We have thus removed all restrictions except the subexponential decay in our approach. A clear picture of the three step strategy emerges: Step 2 and 3 hold under very general conditions and are model independent. The main task of proving the universality is to establish a strong version of the local semicircle law—which can be model dependent. We believe that our method applies to generalized sample covariance matrices as well, but we will not pursue this direction in this paper.

3 Proof of Universality

We now prove the main universality theorem, Theorem 2.2.
Step 1. *Universality for Dyson Brownian Motion:* Under the Assumptions II–IV in the introduction, the universality for the Dyson Brownian Motion was proved in [18]. We recall the statement in the following Theorem.

Theorem 3.1 *[Theorem 2.1 of [18]] Let $\varepsilon > 0$ be the exponent from Assumption III. Suppose that the Assumptions II, III and IV hold for the solution f_t of the forward equation* (2.20) *for all time $t \geq N^{-2\varepsilon}$. Let $E \in \mathbb{R}$ be a point where $\varrho(E) > 0$. Then for any $k \geq 1$ and for any compactly supported continuous test function $O : \mathbb{R}^k \to \mathbb{R}$, we have*

$$\lim_{b\to 0}\lim_{N\to\infty}\sup_{t\geq N^{-2\varepsilon+\delta}} \frac{1}{2b}\int_{E-b}^{E+b} \mathrm{d}E' \int_{\mathbb{R}^k} \mathrm{d}\alpha_1 \ldots \mathrm{d}\alpha_k \, O(\alpha_1,\ldots,\alpha_k) \times \frac{1}{\varrho(E)^k}\Big(p_{t,N}^{(k)} - p_{\mu,N}^{(k)}\Big)\Big(E' + \frac{\alpha_1}{N\varrho(E)},\ldots,E' + \frac{\alpha_k}{N\varrho(E)}\Big) = 0\,. \tag{3.1}$$

Notice that the assumption on the initial entropy is not needed as was remarked in [19].

Step 2 *Universality for Gaussian divisible ensembles:* The Dyson Brownian motion is generated by the matrix flow (2.18). Our task is to determine the initial ensemble H_0 so that the Assumptions II–IV of

Theorem 3.1 can be proved for the flow. The Assumption IV is a direct consequence of the local semicircle law, i.e., Theorem 4.1. The Assumption III will be proved in Proposition 7.1. For the generalized Wigner matrices, the only assumption of Theorem 4.1 and Proposition 7.1 is the subexponential decay property of the distributions of the matrix elements. Since the evolution of the matrix element is given by an Ornstein-Uhlenbeck process, the subexponential property is preserved and we only have to check it for the initial data. We have thus proved the following theorem.

Theorem 3.2 *Suppose that the probability law for the initial matrix H_0 satisfies the assumptions of Theorem 2.2. Then there exists $\varepsilon_0 > 0$ such that for any $t \geq N^{-\varepsilon_0}$, the probability law for the eigenvalues of H_t satisfies the universality equation* (2.17).

Step 3 *Green function comparison theorem:* We have proved the universality for all ensembles with the matrix element at (i, j) distributed by $\sigma_{ij}\xi_t^{ij}$ with

$$\xi_t^{ij} = e^{-t/2}\xi_0^{ij} + (1 - e^{-t})^{1/2}\, \xi_G^{ij}, \tag{3.2}$$

where ξ_G^{ij} are independent Gaussian random variables with mean 0 and variance 1 and $t \sim N^{-\varepsilon}$. In order to prove Theorem 2.2, it remains to approximate all random variables with the subexponential property by ξ_t. The only requirement of ξ_0 is the subexponential decay property and the mean zero and variance one normalization. Our tool is the following Green function comparison theorem from [19]. It implies that the correlation functions of the eigenvalues of two matrix ensembles at a fixed energy are identical up to the scale $1/N$ provided that the first four moments of the matrix elements of these two ensembles are almost identical. Prior to this theorem, it was [28] proved that the joint distribution of individual eigenvalues for Wigner ensembles is the same under the four moment assumption. Tao-Vu's theorem addresses the distribution of individual eigenvalues[1] while Theorem 3.3 compares Green functions (and thus eigenvalues) at a fixed energy.

Theorem 3.3 *Suppose that we have two generalized $N \times N$ Wigner matrices, $H^{(v)}$ and $H^{(w)}$, with matrix elements h_{ij} given by the random variables $N^{-1/2}v_{ij}$ and $N^{-1/2}w_{ij}$, respectively, with v_{ij} and w_{ij} satisfying the uniform subexponential decay condition* (2.11). *Fix a bijective ordering map on the index set of the independent matrix elements,*

$$\phi : \{(i, j) : 1 \leq i \leq j \leq N\} \to \Big\{1, \ldots, \gamma(N)\Big\}, \qquad \gamma(N) := \frac{N(N+1)}{2},$$

and denote by H_γ the generalized Wigner matrix whose matrix elements h_{ij} follow the v-distribution if $\phi(i, j) \leq \gamma$ and they follow the w-distribution otherwise; in particular $H^{(v)} = H_0$ and $H^{(w)} = H_{\gamma(N)}$. Let $\kappa > 0$ be arbitrary and suppose that for any small parameter $\tau > 0$ and for any $y \geq N^{-1+\tau}$ we have the following estimate on the diagonal elements of the resolvent:

$$\mathbb{P}\left(\max_{0\leq\gamma\leq\gamma(N)} \max_{1\leq k\leq N} \max_{|E|\leq 2-\kappa} \left| \left(\frac{1}{H_\gamma - E - iy} \right)_{kk} \right| \leq N^{2\tau} \right) \geq 1 - CN^{-c\log\log N} \tag{3.3}$$

[1] In a recent preprint [31] (appeared after the current preprint was first posted), it was pointed out that if the four moment condition is violated, then the differences between individual eigenvalues of the two ensembles are bigger than the eigenvalue spacing. Thus the four moment condition is also necessary for locating the individual eigenvalues. This is in contrast with the main theme of this paper that gap distribution and correlation functions are even independent of the second moments as long as they are nonzero.

with some constants C, c depending only on τ, κ. Moreover, we assume that the first three moments of v_{ij} and w_{ij} are the same, i.e.

$$\mathbb{E}\bar{v}_{ij}^s v_{ij}^u = \mathbb{E}\bar{w}_{ij}^s w_{ij}^u, \qquad 0 \le s+u \le 3,$$

and the difference between the fourth moments of v_{ij} and w_{ij} is much less than 1, say

$$\left|\mathbb{E}\bar{v}_{ij}^s v_{ij}^{4-s} - \mathbb{E}\bar{w}_{ij}^s w_{ij}^{4-s}\right| \le N^{-\delta}, \qquad s = 0,1,2,3,4, \tag{3.4}$$

for some given $\delta > 0$. Let $\varepsilon > 0$ be arbitrary and choose an η with $N^{-1-\varepsilon} \le \eta \le N^{-1}$. For any sequence of positive integers $k_1, \ldots, k_n$, set complex parameters $z_j^m = E_j^m \pm i\eta$, $j = 1, \ldots k_i$, $m = 1, \ldots, n$ with $|E_j^m| \le 2 - 2\kappa$ and with an arbitrary choice of the $\pm$ signs. Let $G^{(v)}(z) = (H^{(v)} - z)^{-1}$ be the resolvent and let $F(x_1, \ldots, x_n)$ be a function such that for any multi-index $\alpha = (\alpha_1, \ldots, \alpha_n)$ with $1 \le |\alpha| \le 5$ and for any $\varepsilon' > 0$ sufficiently small, we have

$$\max\left\{|\partial^\alpha F(x_1, \ldots, x_n)| : \max_j |x_j| \le N^{\varepsilon'}\right\} \le N^{C_0\varepsilon'} \tag{3.5}$$

and

$$\max\left\{|\partial^\alpha F(x_1, \ldots, x_n)| : \max_j |x_j| \le N^2\right\} \le N^{C_0} \tag{3.6}$$

for some constant C_0.

Then, there is a constant C_1, depending on α, β, $\sum_i k_i$ and C_0 such that for any η with $N^{-1-\varepsilon} \le \eta \le N^{-1}$ and for any choices of the signs in the imaginary part of z_j^m

$$\left|\mathbb{E}F\left(\frac{1}{N^{k_1}} Tr\left[\prod_{j=1}^{k_1} G^{(v)}(z_j^1)\right], \ldots, \frac{1}{N^{k_n}} Tr\left[\prod_{j=1}^{k_n} G^{(v)}(z_j^n)\right]\right) - \mathbb{E}F\left(G^{(v)} \to G^{(w)}\right)\right|$$
$$\le C_1 N^{-1/2 + C_1\varepsilon} + C_1 N^{-\delta + C_1\varepsilon}, \tag{3.7}$$

where in the second term the arguments of F are changed from the Green functions of $H^{(v)}$ to $H^{(w)}$ and all other parameters remain unchanged.

Given this theorem, for any matrix ensemble H whose matrix element at (i,j) are distributed according to $\sigma_{ij}\zeta^{ij}$, we need to find ξ_0^{ij} such that the first four moments of ζ^{ij} and ξ_t^{ij} are almost the same and ξ_0^{ij} has a subexponential decay. Since the real and imaginary parts are i.i.d., it is sufficient to match them individually. This is the content of the following lemma which is stated for real random variables normalized to variance one. With this lemma, we have proved Theorem 2.2. This lemma is essentially the same as Lemma 28 in [28].

Lemma 3.4 *Let m_3 and m_4 be two real numbers such that*

$$m_4 - m_3^2 - 1 \ge 0, \;\; m_4 \le C_2$$

for some positive constant C_2. Let ξ^G be a real Gaussian random variable with mean 0 and variance 1. Then for any sufficient small $\gamma > 0$ (depending on C_2), there exists a real random variable ξ_γ with subexponential decay and independent of ξ^G, such that the first four moments of

$$\xi' = (1-\gamma)^{1/2}\xi_\gamma + \gamma^{1/2}\xi^G$$

are $m_1(\xi') = 0$, $m_2(\xi') = 1$, $m_3(\xi') = m_3$ *and* $m_4(\xi')$, *and*

$$|m_4(\xi') - m_4| \le C\gamma \tag{3.8}$$

for some positive constant C *depending on* C_2.

Proof. It is easy to see by an explicit construction that the following holds:

$$\begin{aligned}&\text{For any given numbers } m_3, m_4, \text{ with } m_4 - m_3^2 - 1 \ge 0 \text{ there is a random}\\ &\text{variable } X \text{ with first four moments } 0, 1, m_3, m_4 \text{ and with subexponential decay.}\end{aligned} \tag{3.9}$$

For any real random variable ζ, independent of ξ^G, and with the first 4 moments being 0, 1, $m_3(\zeta)$ and $m_4(\zeta) < \infty$, the first 4 moments of

$$\zeta' = (1-\gamma)^{1/2}\zeta + \gamma^{1/2}\xi^G$$

are 0, 1,

$$m_3(\zeta') = (1-\gamma)^{3/2} m_3(\zeta) \tag{3.10}$$

and

$$m_4(\zeta') = (1-\gamma)^2 m_4(\zeta) + 6\gamma - 3\gamma^2. \tag{3.11}$$

Using (3.9), we obtain that for any $\gamma > 0$ there exists a real random variable ξ_γ such that the first four moments are 0, 1,

$$m_3(\xi_\gamma) = (1-\gamma)^{-3/2} m_3$$

and

$$m_4(\xi_\gamma) = m_3(\xi_\gamma)^2 + (m_4 - m_3^2).$$

With $m_4 \le C_2$, we have $m_3^2 \le C_2^{3/2}$, thus

$$|m_4(\xi_\gamma) - m_4| \le C\gamma$$

for some positive constant C depending on C_2. Hence with (3.10) and (3.11), we obtain that $\xi' = (1-\gamma)^{1/2}\xi_\gamma + \gamma^{1/2}\xi^G$ satisfies $m_3(\xi') = m_3$ and (3.8). This completes the proof of Lemma 3.4. □

4 Large Deviation of Local Semicircle Law

We first reprove the large deviation of local semicircle law given in [19]. The result of this section is relevant only for $\eta \ge M^{-1}$.

Theorem 4.1 *Assume the* $N \times N$ *random matrix* H *satisfies* (2.1), (2.4), (2.5) *and* (2.11), $\mathbb{E}\, h_{ij} = 0$, *for any* $1 \le i, j \le N$. *Let* $z = E + i\eta$ $(\eta > 0)$ *and let* $\theta(z)$ *be a non-negative function defined by*

$$\theta = \theta(z) := \frac{1}{|1 - m_{sc}(z)^2|} + \frac{1}{\max\big\{\delta_+ \,,\, |\mathfrak{Re}\, m_{sc}^2(z) - 1|\big\}}. \tag{4.1}$$

Let $\kappa \equiv ||E| - 2|$. *Then for all* $z = E + i\eta$ *with*

$$|E| \le 5, \qquad \frac{1}{N} < \eta \le 10, \qquad \sqrt{M\eta} \ge (\log N)^{12+3\alpha}\theta^2(z)(\kappa+\eta)^{1/4} \tag{4.2}$$

we have

$$\mathbb{P}\left\{\max_i |G_{ii}(z) - m_{sc}(z)| \geq (\log N)^{6+2\alpha} \frac{(\kappa+\eta)^{1/4}}{\sqrt{M\eta}}\ \theta(z)\right\} \leq CN^{-c(\log\log N)} \tag{4.3}$$

and

$$\mathbb{P}\left\{\max_{i\neq j} |G_{ij}(z)| \geq (\log N)^{6+2\alpha} \frac{(\kappa+\eta)^{1/4}}{\sqrt{M\eta}}\right\} \leq CN^{-c(\log\log N)} \tag{4.4}$$

for sufficiently large N with positive some constants c and $C > 0$ that depend only α and β in (2.11) *and δ_- in* (2.4) *and* (2.5)*.*

The theorem will be proved at the end of the section after collecting several lemmas. The first lemma describes the behavior of m_{sc} in the various regimes, its proof is elementary calculus. We use the notation $f \sim g$ for two positive functions in some domain D if there is a positive universal constant C such that $C^{-1} \leq f(z)/g(z) \leq C$ holds for all $z \in D$.

Lemma 4.2 *We have for all z with $\mathfrak{Im}\, z > 0$ that*

$$|m_{sc}(z)| = |m_{sc}(z) + z|^{-1} \leq 1. \tag{4.5}$$

From now on, let $z = E + i\eta$ with $|E| \leq 5$ and $\eta > 0$. If $\eta \geq 10$, then we have

$$\mathfrak{Im}\, m_{sc}(z) \sim \eta^{-1}, \qquad |m_{sc}(z)| \sim \eta^{-1}, \qquad |1 - m_{sc}^2(z)| \sim 1, \qquad |1 - \mathfrak{Re}\, m_{sc}^2(z)| \sim 1. \tag{4.6}$$

If $\eta \leq 10$, then we have

$$|m_{sc}(z)| \sim 1, \qquad |1 - m_{sc}^2(z)| \sim \sqrt{\kappa+\eta}. \tag{4.7}$$

For the behavior of $|1 - \mathfrak{Re}\, m_{sc}^2(z)|$ and $\mathfrak{Im}\, m_{sc}(z)$ we distinguish two cases.

Case 1. For $|E| \geq 2$ we have

$$\mathfrak{Im}\, m_{sc}(z) \sim \begin{cases} \frac{\eta}{\sqrt{\kappa+\eta}} & \text{if } \kappa \geq \eta \\ \\ \sqrt{\kappa+\eta} & \text{if } \kappa \leq \eta \end{cases} \tag{4.8}$$

$$|1 - \mathfrak{Re}\, m_{sc}^2(z)| \sim \sqrt{\kappa+\eta}.$$

Case 2. For $|E| \leq 2$ we have

$$\mathfrak{Im}\, m_{sc}(z) \sim \sqrt{\kappa+\eta},$$

$$|1 - \mathfrak{Re}\, m_{sc}^2(z)| \sim \begin{cases} \kappa + \frac{\eta}{\sqrt{\kappa+\eta}} & \text{if } \eta \leq \kappa \\ \\ \sqrt{\kappa+\eta} & \text{if } \kappa \leq \eta \end{cases} \tag{4.9}$$

□

Thus the control function $\theta(z)$ has the following behavior

$$\theta(z) \sim \begin{cases} 1 & \text{if } \ \eta \geq 10, \\ \\ \min\left\{\delta_+^{-1},\ \ \sqrt{\kappa}/\eta,\ \ \kappa^{-1}\right\} & \text{if } \ \eta \leq 10, \quad |E| \leq 2 \ \text{ and } \ \kappa \geq \eta, \\ \\ (\kappa+\eta)^{-1/2} & \text{if } \ \eta \leq 10, \quad \text{and } \ \left\{2 \leq |E| \leq 10 \ \text{ or } \ \kappa \leq \eta\right\}. \end{cases} \tag{4.10}$$

Note that the precise formula (4.1) for $\theta(z)$ is not important, only its asymptotic behavior for small κ, η and δ_+ is relevant. The theorem remains valid if $\theta(z)$ is replaced by $\widetilde{\theta}(z)$ with $\widetilde{\theta}(z) \le C\theta(z)$. In particular, $\theta(z)$ can be chosen to be order one when E is not near the edges of the spectrum. If we are only concerned with the generalized Wigner ensemble (2.6), then by (2.7) we can choose $\theta(z) = (\kappa+\eta)^{-1/2}$ for any $z = E+i\eta$ ($\eta > 0$). For universal Wigner matrices we have $\theta(z) \le C(\kappa+\eta)^{-1}$ for $|z| \le 10$, i.e., using the parameter A introduced in Theorem 2.1, we have

$$\theta(z) \le \frac{C}{(\kappa+\eta)^{A/2}}, \qquad A = 1, 2, \quad |E|, \eta \le 10. \tag{4.11}$$

Based upon these formulas, we also have, for any $z = E + i\eta$ with $\eta > 0$,

$$\mathfrak{Im}\, m_{sc}(z) + \frac{1}{\theta(z)} \le C \min\{1, \sqrt{\kappa+\eta}\}. \tag{4.12}$$

First, we introduce some notations. Recall that $G_{ij} = G_{ij}(z)$ denotes the matrix element

$$G_{ij} = \left(\frac{1}{H-z}\right)_{ij}$$

and

$$m(z) = m_N(z) = \frac{1}{N}\sum_{i=1}^{N} G_{ii}(z).$$

Definition 4.1 *Let $\mathbb{T} = \{k_1, k_2, \ldots, k_t\} \subset \{1, 2, \ldots, N\}$ be an unordered set of $|\mathbb{T}| = t$ elements and let $H^{(\mathbb{T})}$ be the $N-t$ by $N-t$ minor of H after removing the k_i-th $(1 \le i \le t)$ rows and columns. For $\mathbb{T} = \emptyset$, we have $H^{(\emptyset)} = H$. Similarly, we define $\mathbf{a}^{(\ell;\, \mathbb{T})}$ the ℓ-th column with k_i-th $(1 \le i \le t)$ elements removed. Sometimes, we just use the short notation $\mathbf{a}^\ell$=$\mathbf{a}^{(\ell;\, \mathbb{T})}$. For any $\mathbb{T} \subset \{1, 2, \ldots, N\}$ we introduce the following notations:*

$$\begin{aligned}
G^{(\mathbb{T})}_{ij} :=& (H^{(\mathbb{T})} - z)^{-1}(i,j) \\
Z^{(\mathbb{T})}_{ij} :=& \mathbf{a}^i \cdot (H^{(\mathbb{T})} - z)^{-1} \mathbf{a}^j = \sum_{k,l \notin \mathbb{T}} \overline{\mathbf{a}^i_k} G^{(\mathbb{T})}_{k\,l} \mathbf{a}^j_l \\
K^{(\mathbb{T})}_{ij} :=& h_{ij} - z\delta_{ij} - Z^{(\mathbb{T})}_{ij}.
\end{aligned}$$

These quantities depend on z, but we mostly neglect this dependence in the notation.

The following two results were proved in our previous work (Lemma 4.2 and Corollary B.3 of [19]) and they will be our key inputs. We start with the self-consistent perturbation formulas.

Lemma 4.3 *[Self-consistent Perturbation Formulas] Let $\mathbb{T} \subset \{1, 2, \ldots, N\}$. For simplicity, we use the notation $(i\,\mathbb{T})$ for $(\{i\} \cup \mathbb{T})$ and $(ij\,\mathbb{T})$ for $(\{i,j\} \cup \mathbb{T})$. Then we have the following identities:*

1. *For any $i \notin \mathbb{T}$*

$$G^{(\mathbb{T})}_{ii} = (K^{(i\,\mathbb{T})}_{ii})^{-1}. \tag{4.13}$$

2. For $i \neq j$ and $i, j \notin \mathbb{T}$

$$G_{ij}^{(\mathbb{T})} = -G_{jj}^{(\mathbb{T})} G_{ii}^{(j\,\mathbb{T})} K_{ij}^{(ij\,\mathbb{T})} = -G_{ii}^{(\mathbb{T})} G_{jj}^{(i\,\mathbb{T})} K_{ij}^{(ij\,\mathbb{T})}. \tag{4.14}$$

3. For $i \neq j$ and $i, j \notin \mathbb{T}$

$$G_{ii}^{(\mathbb{T})} - G_{ii}^{(j\,\mathbb{T})} = G_{ij}^{(\mathbb{T})} G_{ji}^{(\mathbb{T})} (G_{jj}^{(\mathbb{T})})^{-1}. \tag{4.15}$$

4. For any indices i, j and k that are different and $i, j, k \notin \mathbb{T}$

$$G_{ij}^{(\mathbb{T})} - G_{ij}^{(k\,\mathbb{T})} = G_{ik}^{(\mathbb{T})} G_{kj}^{(\mathbb{T})} (G_{kk}^{(\mathbb{T})})^{-1}. \tag{4.16}$$

Lemma 4.4 *Let a_i $(1 \leq i \leq N)$ be N independent random complex variables with mean zero, variance σ^2 and having the uniform subexponential decay* (2.11). *Let A_i, $B_{ij} \in \mathbb{C}$ $(1 \leq i, j \leq N)$. Then we have that*

$$\mathbb{P}\left\{\left|\sum_{i=1}^N a_i A_i\right| \geq (\log N)^{\frac{3}{2}+\alpha} \sigma \left(\sum_i |A_i|^2\right)^{1/2}\right\} \leq C N^{-\log\log N}, \tag{4.17}$$

$$\mathbb{P}\left\{\left|\sum_{i=1}^N \overline{a}_i B_{ii} a_i - \sum_{i=1}^N \sigma^2 B_{ii}\right| \geq (\log N)^{\frac{3}{2}+2\alpha} \sigma^2 \Big(\sum_{i=1}^N |B_{ii}|^2\Big)^{1/2}\right\} \leq C N^{-\log\log N}, \tag{4.18}$$

$$\mathbb{P}\left\{\left|\sum_{i \neq j} \overline{a}_i B_{ij} a_j\right| \geq (\log N)^{3+2\alpha} \sigma^2 \Big(\sum_{i \neq j} |B_{ij}|^2\Big)^{1/2}\right\} \leq C N^{-\log\log N}, \tag{4.19}$$

for some constants C depending on α and β in (2.11).

We start with determining a system of self-consistent equations for the diagonal matrix elements of the resolvent. We can write G_{ii} as follows,

$$G_{ii} = (K_{ii}^{(i)})^{-1} = \frac{1}{\mathbb{E}_{\mathbf{a}^i} K_{ii}^{(i)} + K_{ii}^{(i)} - \mathbb{E}_{\mathbf{a}^i} K_{ii}^{(i)}},$$

where $\mathbb{E}_{\mathbf{a}^i} = \mathbb{E}_i$ denotes the expectation with respect to the elements in the i-th column of the matrix H, i.e., w.r.t. $\mathbf{a}^i = (h_{1i}, h_{2i}, \ldots, h_{Ni})^t$. Introduce the notations

$$A_i := \sigma_{ii}^2 G_{ii} + \sum_{j \neq i} \sigma_{ij}^2 \frac{G_{ij} G_{ji}}{G_{ii}} \tag{4.20}$$

and

$$Z_i := \sum_{k,l \neq i} \left[\overline{\mathbf{a}_k^i} G_{k\,l}^{(i)} \mathbf{a}_l^i - \mathbb{E}_{\mathbf{a}^i} \overline{\mathbf{a}_k^i} G_{k\,l}^{(i)} \mathbf{a}_l^i\right] = Z_{ii}^{(i)} - \mathbb{E}_i Z_{ii}^{(i)}.$$

Using the fact that $G^{(i)} = (H^{(i)} - z)^{-1}$ is independent of $\mathbf{a}^i$ and $\mathbb{E}_{\mathbf{a}^i} \overline{\mathbf{a}_k^i} \mathbf{a}_l^i = \delta_{k\,l} \sigma_{ik}^2$, we obtain

$$\mathbb{E}_{\mathbf{a}^i} K_{ii}^{(i)} = -z - \sum_{j \neq i} \sigma_{ij}^2 G_{jj}^{(i)}$$

and

$$K_{ii}^{(i)} - \mathbb{E}_{\mathbf{a}^i} K_{ii}^{(i)} = h_{ii} - Z_i.$$

Denote by

$$\Upsilon_i = \Upsilon_i(z) := A_i + \Big(K_{ii}^{(i)} - \mathbb{E}_{\mathbf{a}^i} K_{ii}^{(i)} \Big) = A_i + h_{ii} - Z_i \tag{4.21}$$

and we have the identity

$$G_{ii} = \frac{1}{-z - \sum_j \sigma_{ij}^2 G_{jj} + \Upsilon_i}. \tag{4.22}$$

Let

$$v_i := G_{ii} - m_{sc}, \qquad m := \frac{1}{N}\sum_i G_{ii}, \qquad \bar{v} := \frac{1}{N}\sum_i v_i = \frac{1}{N}\sum_i (G_{ii} - m_{sc}).$$

We will estimate the following key quantities

$$\Lambda_d := \max_k |v_k| = \max_k |G_{kk} - m_{sc}|, \qquad \Lambda_o := \max_{k \neq \ell} |G_{k\ell}|, \tag{4.23}$$

where the subscripts refer to "diagonal" and "offdiagonal" matrix elements. All the quantities defined so far depend on the spectral parameter $z = E + i\eta$, but we will mostly omit this fact from the notation. The real part E will always be kept fixed. For the imaginary part we will use a continuity argument at the end of the proof and then the dependence of $\Lambda_{d,o}$ on η will be indicated.

Both quantities Λ_d and Λ_o will be typically small, eventually we will prove that their size is less than $(M\eta)^{-1/2}$, modulo logarithmic corrections and a factor involving the distance to the edge. We thus define the exceptional event

$$\Omega_\Lambda = \Omega_\Lambda(z) := \Big\{ \Lambda_d(z) + \Lambda_o(z) \geq \frac{(\log N)^{-3/2}}{\theta(z)} \Big\}. \tag{4.24}$$

We will always work in Ω_Λ^c, and, in particular, we will have

$$\Lambda_d(z) + \Lambda_o(z) \leq C(\log N)^{-3/2}$$

since $1/\theta(z) \leq C$ by (4.12). Define the set

$$S := \{ z = E + i\eta \ : \ |E| \leq 5, \quad N^{-1} < \eta \leq 10 \}.$$

We thus have

$$c \leq |G_{ii}(z)| \leq C \qquad \text{in } \Omega_\Lambda^c \tag{4.25}$$

for any $z \in S$ with some universal constant $c > 0$. Here we estimated $\big| |G_{ii}| - |m_{sc}| \big| \leq \Lambda_d$, and we used from (4.6)–(4.7) that $m_{sc}(z)$ satisfies $|m_{sc}(z)| \sim 1$ for $z \in S$.

Thus, a special case of (4.16) or (4.15),

$$G_{k\,l}^{(i)} = G_{k\,l} - \frac{G_{ki}G_{il}}{G_{ii}}, \qquad i \neq l, k,$$

together with (4.25) implies that for any i and with a sufficiently large constant C

$$\max_{k \neq l} |G_{k\,l}^{(i)}| \leq \Lambda_o + C\Lambda_o^2 \leq C\Lambda_o \qquad \text{in } \Omega_\Lambda^c, \tag{4.26}$$

$$C^{-1} \leq |G_{kk}^{(i)}| \leq C, \qquad \text{for all } k \neq i \text{ and in } \Omega_\Lambda^c \tag{4.27}$$

$$|G_{kk}^{(i)} - m_{sc}| \le \Lambda_d + C\Lambda_o^2 \qquad \text{for all } k \neq i \text{ and in } \Omega_\Lambda^c \tag{4.28}$$

and

$$|A_i| \le \frac{C}{M} + C\Lambda_o^2 \qquad \text{in } \Omega_\Lambda^c. \tag{4.29}$$

Here we have used that

$$\Big|\frac{G_{ki}G_{il}}{G_{ii}}\Big| \le c^{-1}\Lambda_o^2 \qquad \text{in } \Omega_\Lambda^c$$

with c being the constant in (4.25) and we also used that $\sum_j \sigma_{ij}^2 = 1$. Similarly, with one more expansion step, we get

$$\max_{ij}\max_{k\neq l}|G_{k\,l}^{(ij)}| \le C\Lambda_o, \qquad \max_{ij}\max_{k}|G_{kk}^{(ij)}| \le C \qquad \text{in } \Omega_\Lambda^c \tag{4.30}$$

and

$$|G_{kk}^{(ij)} - m_{sc}| \le \Lambda_d + C\Lambda_o^2 \qquad \text{for all } k \neq i, j \text{ and in } \Omega_\Lambda^c. \tag{4.31}$$

Using these estimates, the following lemma shows that Z_i and $Z_{ij}^{(ij)}$ are small assuming $\Lambda_d + \Lambda_o$ is small and the h_{ij}'s are not too large. The control parameter for the Z's is $\Phi = \Phi(z)$, defined below (4.32). These bounds hold uniformly in S.

Lemma 4.5 *Denote by*

$$\Phi := \Phi(z) = \frac{\sqrt{\Lambda_d} + \Lambda_o + (\kappa+\eta)^{1/4}}{\sqrt{M\eta}}, \tag{4.32}$$

and define the exceptional events

$$\Omega_1 := \Big\{ \max_{1\le i,j\le N} |h_{ij}| \ge (\log N)^{2\alpha}|\sigma_{ij}| \Big\}$$
$$\Omega_d(z) := \Big\{ \max_i |Z_i(z)| \ge (\log N)^{5+2\alpha}\Phi(z) \Big\}$$
$$\Omega_o(z) := \Big\{ \max_{i\neq j} |Z_{ij}^{(ij)}(z)| \ge (\log N)^{5+2\alpha}\Phi(z) \Big\}$$

and we let

$$\Omega := \Omega_1 \cup \bigcup_{z\in S}\Big[\big(\Omega_d(z)\cup\Omega_o(z)\big) \cap \Omega_\Lambda^c(z) \Big] \tag{4.33}$$

to be the set of all exceptional events. Then we have

$$\mathbb{P}(\Omega) \le CN^{-c(\log\log N)}. \tag{4.34}$$

Proof: Under the assumption of (2.11), we have

$$\mathbb{P}\left(\Omega_1\right) \le CN^{-c\log\log N}, \tag{4.35}$$

therefore we can work on the complement set Ω_1^c. Define the event

$$\widetilde{\Omega}_\Lambda(z) := \Big\{ \Lambda_d(z) + \Lambda_o(z) \ge 2\frac{(\log N)^{-3/2}}{\theta(z)} \Big\}.$$

Notice that the estimates (4.26)–(4.31) also hold on $\widetilde{\Omega}^c_\Lambda$, maybe with different constants C. We now prove that for any fixed $z \in S$, we have

$$\mathbb{P}\Big(\widetilde{\Omega}^c_\Lambda(z) \cap \Big\{\max_i |Z_i(z)| \geq C(\log N)^{2+2\alpha}\Phi(z)\Big\}\Big) \leq CN^{-c\log\log N} \tag{4.36}$$

and

$$\mathbb{P}\Big(\widetilde{\Omega}^c_\Lambda(z) \cap \Big\{\max_{i\neq j} |Z^{(ij)}_{ij}(z)| \geq (\log N)^{4+2\alpha}\Phi(z)\Big\}\Big) \leq CN^{-c\log\log N}. \tag{4.37}$$

To see (4.36), we apply the estimate (4.18) from the large deviation Lemma 4.4, and we obtain that

$$|Z_i| \leq C(\log N)^{\frac{3}{2}+2\alpha}\sqrt{\sum_{k,l\neq i}\Big|\sigma_{ik}G^{(i)}_{k\,l}\sigma_{li}\Big|^2} \tag{4.38}$$

holds with a probability larger than $1 - CN^{-c(\log\log N)}$ for sufficiently large N.

Denote by $u^{(i)}_\alpha$ and $\lambda^{(i)}_\alpha$ ($\alpha = 1, 2, \ldots, N-1$) the eigenvectors and eigenvalues of $H^{(i)}$. Let $u^{(i)}_\alpha(l)$ denote the l-th coordinate of $u^{(i)}_\alpha$. Then, using $\sigma^2_{il} \leq 1/M$ and (4.28), we have

$$\begin{aligned}\sum_{k,l\neq i}\Big|\sigma_{ik}G^{(i)}_{k\,l}\sigma_{li}\Big|^2 &\leq \frac{1}{M}\sum_{k\neq i}\sigma^2_{ik}\left(|G^{(i)}|^2\right)_{kk}\\ &= \frac{1}{M}\sum_{k\neq i}\sigma^2_{ik}\sum_\alpha\frac{|u^{(i)}_\alpha(k)|^2}{|\lambda^{(i)}_\alpha - z|^2} \leq \frac{1}{M}\sum_{k\neq i}\sigma^2_{ik}\frac{\mathfrak{Im}\, G^{(i)}_{kk}(z)}{\eta}\\ &\leq \frac{\Lambda_d + C\Lambda_o^2 + \mathfrak{Im}\, m_{sc}(z)}{M\eta}\\ &\leq C\Phi^2 \qquad \text{in } \widetilde{\Omega}^c_\Lambda.\end{aligned} \tag{4.39}$$

Here we defined $|A|^2 := A^*A$ for any matrix A and we used (4.12) to estimate $\mathfrak{Im}\, m_{sc}(z)$. Together with (4.38) we have proved (4.36) for a fixed z.

For the offdiagonal estimate (4.37), for $i \neq j$, we have from (4.19) that

$$|Z^{(ij)}_{ij}| \leq C(\log N)^{3+2\alpha}\sqrt{\sum_{k,l\neq i,j}\Big|\sigma_{ik}G^{(ij)}_{k\,l}\sigma_{lj}\Big|^2} \tag{4.40}$$

holds with a probability larger than $1 - CN^{-c(\log\log N)}$ for sufficiently large N. Similarly to (4.39), by using (4.31), we get

$$\sum_{k,l\neq i,j}\Big|\sigma_{ik}G^{(ij)}_{k\,l}\sigma_{lj}\Big|^2 \leq C\Phi^2 \qquad \text{in } \widetilde{\Omega}^c_\Lambda.$$

This proves (4.37).

Now we start proving (4.34). First we choose an N^{-10}-net $\mathcal{N}$ in the set S, i.e., a collection of points, $\{z_n\}_{n\in I} \subset S$, such that for any $z \in S$ there is $\widetilde{z} \in \mathcal{N}$ such that $|z - \widetilde{z}| \leq N^{-10}$. The net can be chosen such that $|I| \leq CN^{20}$. Then (4.36) and (4.37) imply that

$$\mathbb{P}\left(\exists\, \widetilde{z} \in \mathcal{N}\,,\ \text{s.t. } \widetilde{\Omega}^c_\Lambda(\widetilde{z}) \text{ holds and} \max_i |Z_i(\widetilde{z})| + \max_{i\neq j}|Z^{(ij)}_{ij}(\widetilde{z})| \geq 2(\log N)^{4+2\alpha}\Phi(\widetilde{z})\right) \leq CN^{-c\log\log N}. \tag{4.41}$$

Now let $z \in S$ be arbitrary and choose $\widetilde{z} \in \mathcal{N}$ such that $|z - \widetilde{z}| \leq N^{-10}$. For any fixed $i \neq j$, we have

$$\left| |Z_{ij}^{(ij)}(z)| - |Z_{ij}^{(ij)}(\widetilde{z})| \right| \leq |z - \widetilde{z}| \max_{\xi \in S} \left| \frac{\partial Z_{ij}^{(ij)}}{\partial z}(\xi) \right|. \tag{4.42}$$

By $\partial Z_{ij}^{(ij)}/\partial z = -\sum_{s,k,l \notin (ij)} \overline{\mathbf{a}_k^i} G_{ks}^{(ij)} G_{sl}^{(ij)} \mathbf{a}_l^j$ and $\max_{ab} |G_{ab}^{(ij)}| \leq \eta^{-1}$, we have

$$\max_{\xi \in S} \left| \frac{\partial Z_{ij}^{(ij)}}{\partial z}(\xi) \right| \leq \frac{(\log N)^{6\alpha}}{M\eta^2} N^3 \leq N^6, \qquad \text{in } \Omega_1^c.$$

In the last inequality, we used the assumption $\eta \geq N^{-1}$. Thus

$$\left| |Z_{ij}^{(ij)}(z)| - |Z_{ij}^{(ij)}(\widetilde{z})| \right| \leq N^{-4} \qquad \text{in } \Omega_1^c.$$

Since $\Phi \geq M^{-1/2}\eta^{-1/4} \geq cN^{-1}$ for $z \in S$, we obtain

$$\left| |Z_{ij}^{(ij)}(z)| - |Z_{ij}^{(ij)}(\widetilde{z})| \right| \leq \Phi(z) \quad \text{in } \Omega_1^c,$$

and exactly in the same way, we have

$$\Big| \, |Z_i(z)| - |Z_i(\widetilde{z})| \, \Big| \leq \Phi(\widetilde{z}) \quad \text{in } \Omega_1^c.$$

Moreover, by estimating $|\partial_z G| \leq N^2$ in S, we see that $\Lambda_d(z)$, $\Lambda_o(z)$, and $\Phi(z)$ are Lipschitz continuous functions in S with a Lipschitz constant bounded by CN^3. Therefore $\Phi(\widetilde{z})$ can be replaced with $\Phi(z)$ in the lower bound on $|Z_{ij}^{(ij)}(\widetilde{z})|$ and $|Z_i(\widetilde{z})|$ obtained from (4.41), and, furthermore, $\Omega_\Lambda^c(z) \subset \widetilde{\Omega}_\Lambda^c(\widetilde{z})$ using a trivial upper bound $\theta(z) \leq N$. Thus we get

$$\mathbb{P}\left(\exists z \in S \text{ s.t. } \Omega_\Lambda^c(z) \text{ and } \Omega_1^c \text{ hold and} \max_i |Z_i(z)| + \max_{i \neq j} |Z_{ij}^{(ij)}(z)| \geq (\log N)^{5+2\alpha}\Phi(\widetilde{z}) \right) \leq CN^{-c\log\log N}.$$

Combining this with (4.35), we obtain (4.34) and thus Lemma 4.5. □

Our goal is to show that $\Lambda_o(z) + \Lambda_d(z)$ is smaller than $(M\eta)^{-1/2}$ (modulo edge and logarithmic corrections) for any $z \in S$ in the event $\Omega^c(z)$. We will use a continuity argument. In Lemma 4.6 we show for any $z \in S$ that if $\Lambda_o(z) + \Lambda_d(z)$ is smaller than $(\log N)^{-3/2}$, then it is actually also smaller than $(M\eta)^{-1/2}$. In Lemma 4.9 we show that this input condition holds at least for $\mathfrak{Im}\, z = \eta = 10$. Then reducing η, we show by a continuity argument that it holds for each $z \in S$.

Lemma 4.6 (Bootstrap) *Let $z = E + i\eta$ and satisfy (4.2), in particular $z \in S$. Recall Λ_d, Λ_o and Ω defined in (4.23) and (4.33). Then we have that, in the event Ω^c, if*

$$\Lambda_o(z) + \Lambda_d(z) \leq \frac{(\log N)^{-3/2}}{\theta(z)}, \tag{4.43}$$

then we have

$$\Lambda_o(z) + \Lambda_d(z) \leq (\log N)^{6+2\alpha} \frac{(\kappa + \eta)^{1/4}}{\sqrt{M\eta}} \theta(z) \tag{4.44}$$

and we also have a stronger bound for the off-diagonal terms:

$$\Lambda_o(z) \le (\log N)^{5+2\alpha} \frac{(\kappa+\eta)^{1/4}}{\sqrt{M\eta}}. \tag{4.45}$$

Proof of Lemma 4.6. First note that condition (4.43) is equivalent assuming the event $\Omega_\Lambda^c(z)$ and we have

$$\Omega^c \cap \Omega_\Lambda^c(z) \subset \Omega_d^c(z) \cup \Omega_o^c(z), \tag{4.46}$$

so the event $\Omega_d^c(z) \cup \Omega_o^c(z)$ holds. We recall from (4.12) that

$$\frac{1}{\theta(z)} \le C\sqrt{\kappa+\eta} \le C, \qquad z \in S. \tag{4.47}$$

With the assumption (4.43) we have (see (4.25), (4.27))

$$c \le |G_{ii}| \le C, \qquad c \le |G_{jj}^{(i)}| \le C \tag{4.48}$$

and by (4.47)

$$\Lambda_d(z) + \Lambda_o(z) \le \frac{C\sqrt{\kappa+\eta}}{(\log N)^{3/2}} \le \frac{C}{(\log N)^{3/2}} \tag{4.49}$$

and thus, by (4.2) and (4.47),

$$\frac{(\kappa+\eta)^{1/4}}{\sqrt{M\eta}} \le \Phi(z) \le C\frac{(\kappa+\eta)^{1/4}}{\sqrt{M\eta}} \le C(\log N)^{-12-3\alpha}. \tag{4.50}$$

We first estimate the offdiagonal term G_{ij}. From (4.14) we have

$$|G_{ij}| = |G_{ii}||G_{jj}^{(i)}||K_{ij}^{(ij)}| \le C\left(|h_{ij}| + |Z_{ij}^{(ij)}|\right), \qquad i \ne j, \tag{4.51}$$

where we used (4.48).

By the remark after (4.46) we have

$$|G_{ij}| \le \frac{C(\log N)^{2\alpha}}{\sqrt{M}} + C(\log N)^{5+2\alpha}\Phi \le C(\log N)^{5+2\alpha}\Phi,$$

where we used (4.50) to show that the first term can be absorbed into the second. From the second inequality in (4.50) we also have

$$\Lambda_o = \max_{i\ne j}|G_{ij}| \le C(\log N)^{5+2\alpha}\frac{(\kappa+\eta)^{1/4}}{\sqrt{M\eta}}. \tag{4.52}$$

This proves the estimate (4.45). Using (4.47), we also see that (4.44) holds for the summand Λ_o.

Now we estimate the diagonal terms. Recalling $\Upsilon_i = A_i + h_{ii} - Z_i$ from (4.21), with (4.29), (4.50), (4.52) we have,

$$\Upsilon = \Upsilon(z) := \max_i |\Upsilon_i(z)| \le C\frac{(\log N)^{2\alpha}}{\sqrt{M}} + C(\log N)^{5+2\alpha}\Phi \qquad \text{in } \Omega^c \cap \Omega_\Lambda^c(z). \tag{4.53}$$

Again, the first term can be absorbed into the second, so we have proved

$$\Upsilon \leq (\log N)^{5+2\alpha}\Phi \leq C(\log N)^{-6} \qquad \text{in } \Omega^c \cap \Omega_\Lambda^c(z). \tag{4.54}$$

In the last step we used (4.50).

From (4.22) we have the identity

$$v_i = G_{ii} - m_{sc} = \frac{1}{-z - m_{sc} - \Big(\sum_j \sigma_{ij}^2 v_j - \Upsilon_i\Big)} - m_{sc}. \tag{4.55}$$

Using $(m_{sc}+z) = -m_{sc}^{-1}$, and the fact that $|m_{sc}+z| \geq 1$, so with $\Lambda_d + \Upsilon \leq \frac{1}{10}|m_{sc}+z|$ (see in (4.49) and (4.54)), we can expand (4.55) as

$$v_i = m_{sc}^2 \cdot \Big(\sum_j \sigma_{ij}^2 v_j - \Upsilon_i\Big) + O\Big(\sum_j \sigma_{ij}^2 v_j - \Upsilon_i\Big)^2 = m_{sc}^2 \cdot \Big(\sum_j \sigma_{ij}^2 v_j - \Upsilon_i\Big) + O\Big((\Lambda_d + \Upsilon)^2\Big). \tag{4.56}$$

Summing up this formula for all i and recalling the definition $\bar{v} \equiv \frac{1}{N}\sum_i v_i = m - m_{sc}$ yield

$$\bar{v} = m_{sc}^2\,\bar{v} - \frac{m_{sc}^2}{N}\sum_i \Upsilon_i + O\Big((\Lambda_d + \Upsilon)^2\Big).$$

Introducing the notations $\zeta := m_{sc}^2(z)$, $\overline{\Upsilon} := \frac{1}{N}\sum_i \Upsilon_i$ for simplicity, we have (using $\Lambda_d \leq 1$)

$$\bar{v} = -\frac{\zeta}{1-\zeta}\overline{\Upsilon} + O\Big(\frac{\zeta}{1-\zeta}(\Lambda_d + \Upsilon)^2\Big) = O\left(\left|\frac{\zeta}{1-\zeta}\right|\left(\Lambda_d^2 + \Upsilon\right)\right). \tag{4.57}$$

Recall that Σ denotes the matrix of covariances, $\Sigma_{ij} = \sigma_{ij}^2$, and we know that 1 is a simple eigenvalue with the constant vector $\mathbf{e} = N^{-1/2}(1,1,\ldots,1)$ as the eigenvector. Let $Q := I - |\mathbf{e}\rangle\langle\mathbf{e}|$ be the projection onto the orthogonal complement of $\mathbf{e}$, note that Σ and Q commute. Let $\|\cdot\|_{\infty\to\infty}$ denote the $\ell^\infty \to \ell^\infty$ matrix norm. With these notations, (4.56) can be written as

$$v_i - \bar{v} = \zeta\sum_j \Sigma_{ij}(v_j - \bar{v}) - \zeta\Big(\Upsilon_i - \overline{\Upsilon}\Big) + O\left(|\zeta|\left(\Lambda_d^2 + \Upsilon\right)\right) + O\Big((\Lambda_d + \Upsilon)^2\Big),$$

and the error terms for each i sums up to zero. Therefore, with $\Upsilon \leq 1$, we have

$$\begin{aligned} v_i - \bar{v} =& -\sum_j \Big(\frac{\zeta}{1-\zeta\Sigma}\Big)_{ij}\left(\Upsilon_j - \overline{\Upsilon}\right) + O\left(\Big\|\frac{\zeta Q}{1-\zeta\Sigma}\Big\|_{\infty\to\infty}(\Lambda_d^2 + \Upsilon)\right) \\ =& \Big\|\frac{\zeta Q}{1-\zeta\Sigma}\Big\|_{\infty\to\infty} O\big(\Lambda_d^2 + \Upsilon\big). \end{aligned} \tag{4.58}$$

Combining (4.57) with (4.58), we have

$$\max_i |v_i| \leq C\left(\Big\|\frac{\zeta Q}{1-\zeta\Sigma}\Big\|_{\infty\to\infty} + \left|\frac{\zeta}{1-\zeta}\right|\right)(\Lambda_d^2 + \Upsilon). \tag{4.59}$$

To estimate the norm of the resolvent, we recall the following elementary lemma (Lemma 5.3 in [19]).

Lemma 4.7 *Let $\delta_- > 0$ be a given constant. Then there exist small real numbers $\tau \geq 0$ and $c_1 > 0$, depending only on δ_-, such that for any positive number δ_+, we have*

$$\max_{x\in[-1+\delta_-,1-\delta_+]} \left\{ \left|\tau + x\, m_{sc}^2(z)\right|^2 \right\} \leq (1 - c_1\, q(z))\,(1+\tau)^2 \tag{4.60}$$

with

$$q(z) := \max\{\delta_+, |1 - \mathfrak{Re}\; m_{sc}^2(z)|\}. \tag{4.61}$$

□

Lemma 4.8 *Suppose that Σ satisfies* (2.4)*, i.e., Spec($Q\Sigma$) $\subset [-1+\delta_-, 1-\delta_+]$. Then we have*

$$\Big\| \frac{Q}{1 - m_{sc}^2(z)\Sigma} \Big\|_{\infty\to\infty} \leq \frac{C(\delta_-)\log N}{q(z)} \tag{4.62}$$

with some constant $C(\delta_-)$ depending on δ_- and with q defined in (4.61)

Proof: Let $\|\cdot\|$ denote the usual $\ell^2 \to \ell^2$ matrix norm and introduce $\zeta = m_{sc}^2(z)$. Rewrite

$$\Big\| \frac{Q}{1-\zeta\Sigma} \Big\| = \frac{1}{1+\tau} \Big\| \frac{Q}{1 - \frac{\zeta\Sigma+\tau}{1+\tau}} \Big\|$$

with τ given in (4.60). By (4.60), we have

$$\Big\| \frac{\zeta\Sigma+\tau}{1+\tau} Q \Big\| \leq \sup_{x\in[-1+\delta_-,1-\delta_+]} \Big| \frac{\zeta x + \tau}{1+\tau} \Big| \leq (1 - c_1 q(z))^{1/2}.$$

To estimate the $\ell^\infty \to \ell^\infty$ norm of this matrix, recall that $|\zeta| = |m_{sc}|^2 \leq 1$ and $\sum_j |\Sigma_{ij}| = \sum_j \Sigma_{ij} = \sum_j \sigma_{ij}^2 = 1$. Thus we have

$$\Big\| \frac{\zeta\Sigma+\tau}{1+\tau} Q \Big\|_{\infty\to\infty} = \max_i \sum_j \Big| \Big(\frac{\zeta\Sigma+\tau}{1+\tau} \Big)_{ij} \Big| \leq \frac{1}{1+\tau} \max_i \sum_j |\zeta\Sigma_{ij} + \tau\delta_{ij}| \leq \frac{|\zeta|+\tau}{1+\tau} \leq 1.$$

To see (4.62), we can expand

$$\begin{aligned} \Big\| \frac{1}{1 - \frac{\zeta\Sigma+\tau}{1+\tau}} Q \Big\|_{\infty\to\infty} &\leq \sum_{n<n_0} \Big\| \frac{\zeta\Sigma+\tau}{1+\tau} Q \Big\|_{\infty\to\infty}^n + \sum_{n\geq n_0} \Big\| \Big(\frac{\zeta\Sigma+\tau}{1+\tau} \Big)^n Q \Big\|_{\infty\to\infty} \\ &\leq n_0 + \sqrt{N} \sum_{n\geq n_0} \Big\| \Big(\frac{\zeta\Sigma+\tau}{1+\tau} \Big)^n Q \Big\| = n_0 + \sqrt{N} \sum_{n\geq n_0} (1 - c_1 q(z))^{n/2} \\ &= n_0 + C\sqrt{N} \frac{(1 - c_1 q(z))^{n_0/2}}{q(z)} \leq \frac{C\log N}{q(z)}. \end{aligned}$$

Choosing $n_0 = C\log N / q(z)$ with a large C, we have proved the Lemma. □

We now return to the proof of Lemma 4.6, recall that we are in the set $\Omega^c \cap \Omega^c_\Lambda(z)$. First, inserting (4.5) and (4.62) into (4.59), and using $1/q \leq \theta$, we obtain

$$\Lambda_d = \max_i |v_i| \leq C\theta(z)(\Lambda_d^2 + \Upsilon)\log N.$$

By the assumption (4.43), we have $C\theta(z)\Lambda_d \log N \leq 1/2$, for large enough N, therefore we get

$$\Lambda_d \leq C\theta(z)\Upsilon \log N.$$

Using the bound on Υ in (4.54) and (4.50), we obtain

$$\Lambda_d \leq C\theta(z)(\log N)^{6+2\alpha}\frac{(\kappa+\eta)^{1/4}}{\sqrt{M\eta}},$$

which, together with (4.52), completes the proof of (4.44). □

Lemma 4.9 (Initial step) *Define*

$$\Omega_H := \{\|H\| \geq 3\},$$

recall the definitions of Ω_1*,* Ω_d *and* Ω_o *from* (4.33) *and define*

$$\widehat{\Omega} := \Omega_H \cup \Omega_1 \cup \bigcup \Big\{ \Omega_o(z) \cup \Omega_d(z) \;:\; z = E + 10i, |E| \leq 5 \Big\}. \tag{4.63}$$

Then we have

$$\mathbb{P}(\widehat{\Omega}) \leq CN^{-c\log\log N}. \tag{4.64}$$

Furthermore, in the set $\widehat{\Omega}^c$ *we have*

$$\Lambda_o(z) + \Lambda_d(z) \leq \frac{(\log N)^{-3/2}}{\theta(z)} \tag{4.65}$$

for $z = E + 10i$*,* $|E| \leq 5$*.*

Proof. The exceptional event Ω_H is controlled by Lemma 7.2 of [19]. For convenience, we will recall this result in Lemma 6.2, Eq. (6.11), and we note that the condition of this lemma, $M \geq (\log N)^9$, is implied by (4.2) and (4.12)). Thus we have $\mathbb{P}(\Omega_H) \leq CN^{-c(\log\log N)}$.

Denote by u_α and λ_α the eigenvectors and eigenvalues of H. On the set Ω_H^c all eigenvalues are bounded, $|\lambda_\alpha| \leq 3$. In this set we have, with $|E| \leq 5$,

$$\mathfrak{Im}\, G_{kk} = \eta \sum_\alpha \frac{|u_\alpha(k)|^2}{(\lambda_\alpha - E)^2 + \eta^2} \geq \frac{c}{\eta}\sum_\alpha |u_\alpha(k)|^2 = \frac{c}{\eta} \tag{4.66}$$

with some positive constant $c > 0$. We also have the upper bound $|G_{kk}| \leq \eta^{-1}$ and $\Lambda_o + \Lambda_d \leq C/\eta$. In particular, for $\eta = 10$, we have

$$c \leq |G_{kk}| \leq C, \qquad \text{in } \Omega_H^c, \tag{4.67}$$

with some positive constants. Inspecting the proof of Lemma 4.5, notice that the restriction to the set Ω_Λ^c was used only to obtain the estimate (4.25). Once this estimate is obtained independently, as in (4.67) in the set Ω_H^c, all the estimates (4.26)–(4.31) hold and these are the necessary inputs for Lemma 4.5. Thus, following the proof of (4.36)–(4.37), and replacing Ω_Λ^c with Ω_H^c, we obtain that $\mathbb{P}\big\{\Omega_H^c \cap (\Omega_o(z) \cup \Omega_d(z))\big\} \leq CN^{-c\log\log N}$ for each fixed $z = E + 10i$, $|E| \leq 5$. Finally, this estimate can be extended to hold simultaneously for all $z = E + 10i$, $|E| \leq 5$ using an N^{-10}-net as for the proof of (4.34). This proves (4.64).

Similarly, the argument (4.51)–(4.52) shows that in the set $\widehat{\Omega}^c$, we have

$$\Lambda_o(z) \leq \frac{C(\log N)^{5+2\alpha}}{\sqrt{M}}, \qquad z = E + 10i, \tag{4.68}$$

and the argument (4.53)–(4.54) guarantees that

$$\Upsilon(z) \leq \frac{C(\log N)^{5+2\alpha}}{\sqrt{M}}, \qquad z = E + 10i, \tag{4.69}$$

in $\widehat{\Omega}^c$. Finally, to control Λ_d, we use that from the self consistent equation (4.55) and the definition of m_{sc}, we have

$$v_n = \frac{\sum_i \sigma_{ni}^2 v_i + O(\Upsilon)}{(z + m_{sc} + \sum_i \sigma_{ni}^2 v_i + O(\Upsilon))(z + m_{sc})}, \;\; 1 \leq n \leq N. \tag{4.70}$$

For $\eta = 10$, with (2.9), we have $|z + m_{sc}(z)| > 2$. Using $|G_{ii}| \leq \eta^{-1} = \frac{1}{10}$ and $|m_{sc}| \leq \eta^{-1} = \frac{1}{10}$, we obtain

$$|v_i| \leq 2/\eta \leq 1/5, \qquad 1 \leq i \leq N. \tag{4.71}$$

Using (4.69), together with $|z + m_{sc}(z)| > 2$ and (4.71), we obtain that the absolute value of the r.h.s of (4.70) is less than

$$\frac{\sup_i |v_i|}{|z + m_{sc}(z)| - \sup_i |v_i|} + O(\Upsilon). \tag{4.72}$$

Taking the absolute value of (4.70) and maximizing over n, we have

$$\Lambda_d = \sup_n |v_n| \leq \frac{\Lambda_d}{|z + m_{sc}| - \Lambda_d} + O(\Upsilon). \tag{4.73}$$

Since the denominator satisfies $|z + m_{sc}(z)| - \sup_i |v_i| \geq 2 - 1/5$,

$$\Lambda_d \leq C\Upsilon \tag{4.74}$$

follows from the last equation. Combining it with (4.68) and (4.69), we obtain (4.65), and this completes the proof of Lemma 4.9. □

Proof of Theorem 4.1. Lemma 4.6 states that, in the event Ω^c, if $\Lambda_d(z) + \Lambda_o(z) \leq R(z)$ then $\Lambda_d(z) + \Lambda_o(z) \leq S(z)$ with

$$R(z) := (\log N)^{-3/2}(\theta(z))^{-1}, \qquad S(z) := (\log N)^{6+2\alpha} \frac{(\kappa + \eta)^{1/4}}{\sqrt{M\eta}} \theta(z).$$

By assumption (4.2) of Theorem 4.1, we have $S(z) < R(z)$ for any $z \in S$ and these functions are continuous. Lemma 4.9 states that in the set $\widehat{\Omega}^c$ the bound $\Lambda_d(z) + \Lambda_o(z) \leq R(z)$ holds for $\eta = 10$.

Thus by a continuity argument, $\Lambda_d(z) + \Lambda_o(z) \leq S(z)$ in the set $\Omega^c \cap \widehat{\Omega}^c$ as long as the condition (4.2) is satisfied. Finally, once $\Lambda_o(z) \leq S(z)$ is proven, we can use $S(z) \leq R(z)$ (in the domain D) and Lemma 4.6 once more to conclude the stronger bound on $\Lambda_o(z)$. This proves Theorem 4.1.

We record that combining the bound on Λ_d, Λ_o with (4.54), we also proved that under the assumption (4.2) we have

$$\Lambda_d(z) + \Lambda_o(z) + \Upsilon(z) \leq C(\log N)^{16+4\alpha} \frac{(\kappa + \eta)^{1/4}}{\sqrt{M\eta}} \theta(z) \qquad \text{in } \Omega^c \cap \widehat{\Omega}^c. \tag{4.75}$$

□

5 Local semicircle law

In this section we strengthen the estimate of Theorem 4.1 for the Stieltjes transform $m(z) = \frac{1}{N}\sum_i G_{ii}$. The key improvement is that $|m - m_{sc}|$ will be estimated with a precision $(M\eta)^{-1}$ while the $|G_{ii} - m_{sc}|$ was controlled by a precision $(M\eta)^{-1/2}$ only (modulo logarithmic terms and terms expressing the deterioration of the estimate near the edge).

Theorem 5.1 *Assume the conditions of Theorem 4.1 and recall the notations $\kappa = \kappa_E := \big||E| - 2\big|$ and $\theta(z)$ from* (4.1)*. Define the domain*

$$D^* := \Big\{ z = E + i\eta \in \mathbb{C} \ : \ |E| \le 5, \ \frac{1}{N} \le \eta \le 10 \ , \quad M\eta \ge (\log N)^{24+6\alpha}\theta^4(z)(\kappa+\eta)^{1/2} \Big\}. \tag{5.1}$$

Then for any $\varepsilon > 0$ and $K > 0$ there exists a constant $C = C(\varepsilon, K)$ such that

$$\mathbb{P}\Big(\bigcup_{z \in D^*} \Big\{ |m(z) - m_{sc}(z)| \ge \frac{N^\varepsilon \theta^2(z)}{M\eta} \Big\} \Big) \le \frac{C(\varepsilon, K)}{N^K}. \tag{5.2}$$

Proof of Theorem 5.1. We will work in the set $\Omega^c \cap \widehat{\Omega}^c$, which has almost full probability by (4.34) and (4.64). Note that the set D^* is included in the domain defined by (4.2), therefore we can use the estimates from Section 4.

As in (4.57), where $\bar{v} = m(z) - m_{sc}(z)$, we have that

$$m - m_{sc} = -\frac{\zeta}{1-\zeta}\frac{1}{N}\sum_i \Upsilon_i + O\Big(\frac{\zeta}{1-\zeta}(\Lambda_d + \Upsilon)^2\Big)$$

holds with a very high probability. Recall that $\zeta = m_{sc}^2(z)$ and we mostly omit the argument z from the notations. The quantities Λ_d, Υ_i and Υ were defined in (4.23), (4.21) and (4.53). Then with (4.75) we have

$$m(z) - m_{sc}(z) = O\left(\frac{\zeta}{1-\zeta}\,\frac{1}{N}\sum_j \Upsilon_j \right) + O\Big(\frac{N^\varepsilon}{|1-\zeta|}\,\frac{\theta^2(z)\sqrt{\kappa+\eta}}{M\eta} \Big)$$

holds with a very high probability for any small $\varepsilon > 0$. Recall that $\Upsilon_i = A_i + h_{ii} - Z_i$. We have, from (4.20), (4.25) and $\sigma_{ij}^2 \le M^{-1}$,

$$A_j \le \frac{C}{M} + C\Lambda_o^2 \le CN^\varepsilon \frac{\theta^2\sqrt{\kappa+\eta}}{M\eta},$$

where we used (4.75) to bound Λ_o and (4.47) to control the C/M term.

We thus obtain that

$$m - m_{sc} = O\left(\frac{\zeta}{1-\zeta}\left(\frac{1}{N}\sum_i Z_i - \frac{1}{N}\sum_i h_{ii} \right) \right) + O\Big(\frac{N^\varepsilon}{|1-\zeta|}\frac{\theta^2\sqrt{\kappa+\eta}}{M\eta} \Big) \tag{5.3}$$

holds with a very high probability. Since h_{ii}'s are independent, applying the first estimate in the large deviation Lemma 4.4, we have

$$\mathbb{P}\left(\Big| \frac{1}{N}\sum_i h_{ii} \Big| \ge (\log N)^{3/2+\alpha}\frac{1}{\sqrt{MN}} \right) \le CN^{-c\log\log N}. \tag{5.4}$$

On the complement event, the estimate $(\log N)^{3/2+\alpha}(MN)^{-1/2}$ can be included in the last error term in (5.3). It only remains to bound

$$\frac{1}{N}\sum_{i=1}^{N} Z_i,$$

whose moment is bounded in the next lemma which will be proved in Sections 8 and 9.

Lemma 5.2 *For fixed z in domain D^* (5.1) and any even number p, we have*

$$\mathbb{E}\left|\frac{1}{N}\sum_{i=1}^{N} Z_i\right|^p \le C_p\left((\log N)^{3+2\alpha}X^2\right)^p \tag{5.5}$$

for sufficiently large N, where

$$X = X(z) := (\log N)^{10+2\alpha}\frac{(\kappa+\eta)^{1/4}}{\sqrt{M\eta}}, \qquad z = E+i\eta, \quad \kappa = \big||E|-2\big|. \tag{5.6}$$

Using Lemma 5.2, we have that for any $\varepsilon > 0$ and $K > 0$,

$$\mathbb{P}\left(\frac{1}{N}\left|\sum_{i=1}^{N} Z_i\right| \ge N^{\varepsilon}\frac{\sqrt{\kappa+\eta}}{M\eta}\right) \le N^{-K}$$

for sufficiently large N. Combining this with (5.4) and (5.3) and noting that $|1-\zeta| \sim \sqrt{\kappa+\eta}$, see (4.7), we obtain (5.2) and complete the proof of Theorem 5.1.

□

6 Empirical counting function

In this section we translate the information on the Stieltjes transform obtained in Theorem 5.1 to an asymptotic on the empirical counting function. The main ingredient for the first step is the following lemma based upon the Helffer-Sjöstrand formula. We will formulate this lemma for general signed measures, but we will apply it to the Stieltjes transform $m^\Delta = m - m_{sc}$ of the difference between the empirical density and the semicircle law. A similar statement was already proven in Lemma B.1 in [15] and Lemma 7.7 in [19].

Lemma 6.1 *Let ϱ^Δ be a signed measure on the real line with supp $\varrho^\Delta \subset [-K,K]$ for some fixed constant $K \ge 4$. For any $E_1, E_2 \in [-3,3]$ and $\eta \in (0,1/2]$ we define $f(\lambda) = f_{E_1,E_2,\eta}(\lambda)$ to be a characteristic function of $[E_1,E_2]$ smoothed on scale η, i.e., $f \equiv 1$ on $[E_1,E_2]$, $f \equiv 0$ on $\mathbb{R}\setminus[E_1-\eta, E_2+\eta]$ and $|f'| \le C\eta^{-1}$, $|f''| \le C\eta^{-2}$. For any $x \in \mathbb{R}$, set $\kappa_x := \big||x|-2\big|$. Let m^Δ be the Stieltjes transform of ϱ^Δ. Suppose for some positive U, and non-negative constant A we have*

$$|m^\Delta(x+iy)| \le \frac{CU}{y(\kappa_x+y)^A} \qquad \text{for} \qquad 1 \ge y > 0, \quad |x| \le K+1, \tag{6.1}$$

and in case of $A > 0$ we additionally assume $\eta \le \frac{1}{2}\min\{\kappa_{E_1}, \kappa_{E_2}\}$. Then

$$\left|\int f_{E_1,E_2,\eta}(\lambda)\varrho^\Delta(\lambda)\mathrm{d}\lambda\right| \le \frac{CU|\log\eta|}{\big[\min\{\kappa_{E_1},\kappa_{E_2}\}\big]^A} \tag{6.2}$$

with some constant C depending on K and A.

Proof of Lemma 6.1. For simplicity, we drop the Δ superscript in the proof. Analogously to (B.13), (B.14) and (B.15) in [15] we obtain that (with $f = f_{E_1,E_2,\eta}$)

$$\begin{aligned}\left|\int f(\lambda)\varrho(\lambda)\mathrm{d}\lambda\right| \le \quad & C\int_{\mathbb{R}^2}(|f(x)| + |y||f'(x)|)|\chi'(y)||m(x+iy)|\mathrm{d}x\mathrm{d}y \\ & +C\left|\int_{|y|\le\eta}\int yf''(x)\chi(y)\mathfrak{Im}\, m(x+iy)\mathrm{d}x\mathrm{d}y\right| \\ & +C\left|\int_{|y|\ge\eta}\int_{\mathbb{R}} yf''(x)\chi(y)\mathfrak{Im}\, m(x+iy)\mathrm{d}x\mathrm{d}y\right|,\end{aligned} \tag{6.3}$$

where $\chi(y)$ is a smooth cutoff function with support in $[-1, 1]$, with $\chi(y) = 1$ for $|y| \le 1/2$ and with bounded derivatives. The first term is estimated by, with (6.1),

$$\int_{\mathbb{R}^2}(|f(x)| + |y||f'(x)|)|\chi'(y)||m(x+iy)|\mathrm{d}x\mathrm{d}y \le CU. \tag{6.4}$$

For the second term in r.h.s of (6.3) we use that from (6.1) it follows for any $1 \ge y > 0$ that

$$y|\mathfrak{Im}\, m(x+iy)| \le \frac{CU}{(\kappa_x + y)^A}. \tag{6.5}$$

With $|f''| \le C\eta^{-2}$ and

$$\mathrm{supp} f'(x) \subset \{|x - E_1| \le \eta\} \cup \{|x - E_2| \le \eta\}, \tag{6.6}$$

we get

$$\text{second term in r.h.s of (6.3)} \le \frac{CU}{\big[\min\{\kappa_{E_1}, \kappa_{E_2}\}\big]^A}.$$

As in (B.17) and (B.19) in [15], we integrate the third term in (6.3) by parts first in x, then in y. Then we bound it with an absolute value by

$$C\int_{|x|\le K+1}\eta|f'(x)||\mathfrak{Re}\, m(x+i\eta)|\mathrm{d}x + C\int_{\mathbb{R}^2}|f'(x)\chi'(y)\mathfrak{Re}\, m(x+iy)| + \frac{C}{\eta}\int_{\eta\le y\le 1}\int_{|x-E|\le\eta}|\mathfrak{Re}\, m(x+iy)|\mathrm{d}x\mathrm{d}y. \tag{6.7}$$

The second term is bounded in (6.4). By using (6.1) and (6.6) in the first term and (6.1) in the third, we have

$$\begin{aligned}(6.7) \le & \frac{CU}{\big[\min\{\kappa_{E_1}, \kappa_{E_2}\}\big]^A} + CU + CU\eta^{-1}\sum_{k=1,2}\int_{|x-E_k|\le\eta}\mathrm{d}x\int_{\eta\le y\le 1}\frac{1}{y(\kappa_x + y)^A}\mathrm{d}y \\ \le & \frac{CU|\log\eta|}{\big[\min\{\kappa_{E_1}, \kappa_{E_2}\}\big]^A}. \qquad \square\end{aligned}$$

Let $\lambda_1 \le \lambda_2 \le \ldots \le \lambda_N$ be the ordered eigenvalues of a universal Wigner matrix. We define the *normalized empirical counting function* by

$$\mathfrak{n}(E) := \frac{1}{N}\#\{\lambda_j \le E\} \tag{6.8}$$

and the *averaged counting function* by

$$n(E) = \frac{1}{N}\mathbb{E}\#[\lambda_j \le E]. \tag{6.9}$$

Finally, let

$$n_{sc}(E) := \int_{-\infty}^{E} \varrho_{sc}(x)\mathrm{d}x \tag{6.10}$$

be the distribution function of the semicircle law which is very close to the counting function of γ's, $n^\gamma(E) := \frac{1}{N}\#[\gamma_j \le E]$.

We will need some control on the spectral edge, we recall the Lemma 7.2 from [19].

Lemma 6.2 *(1) Let the universal Wigner matrix H satisfy* (2.1)*,* (2.2) *and* (2.11) *with $M \ge (\log N)^9$. Then we have*

$$n(-3) \le CN^{-c\log\log N} \;\text{ and }\; n(3) \ge 1 - CN^{-c\log\log N}. \tag{6.11}$$

(2) Let H be a generalized Wigner matrix with subexponential decay, i.e., (2.1)*,* (2.2)*,* (2.6) *and* (2.11) *hold. Then*

$$n(-2-N^{-1/6+\varepsilon}) \le Ce^{-N^{\varepsilon'}} \;\text{ and }\; n(2+N^{-1/6+\varepsilon}) \ge 1 - Ce^{-N^{\varepsilon'}}, \tag{6.12}$$

for any small $\varepsilon > 0$ with an $\varepsilon' > 0$ depending on ε. Furthermore, for $K \ge 3$,

$$n(-K) \le e^{-N^{\varepsilon}\log K} \;\text{ and }\; n(K) \ge 1 - e^{-N^{\varepsilon}\log K}, \tag{6.13}$$

for some $\varepsilon > 0$.

With these preliminary lemmas, we have the following theorem that we state for universal Wigner matrices and for their subclass, the generalized Wigner matrices in parallel.

Theorem 6.3 *Let $A = 2$ for universal Wigner matrices and $A = 1$ for generalized Wigner matrices. Suppose that the universal Wigner matrix ensemble satisfies* (2.1)*,* (2.2) *and* (2.11) *with $M \ge (\log N)^{24+6\alpha}$ and the generalized Wigner matrix ensemble satisfies* (2.1)*,* (2.2)*,* (2.6) *and* (2.11)*. We recall $M = N$ in the latter case. Then for any $\varepsilon > 0$ and $K \ge 1$ there exists a constant $C(\varepsilon, K)$ such that*

$$\mathbb{P}\Big\{ \sup_{|E|\le 3} \big|\mathfrak{n}(E) - n_{sc}(E)\big|\,[\kappa_E]^A \le \frac{CN^\varepsilon}{M}\Big\} \ge 1 - \frac{C(\varepsilon,K)}{N^K},$$

where the $\mathfrak{n}(E)$ and $n_{sc}(E)$ were defined in (6.8) *and* (6.10) *and $\kappa_E = \big||E|-2\big|$.*

Proof. For definiteness, we will consider the case of generalized Wigner matrices, i.e., $A = 1$. In this case $M = N$, $\delta_+ \ge C_{inf} > 0$ (see (2.7)) and thus $\theta(z) \le C(\kappa+\eta)^{-1/2}$ for $|z| \le 10$, see (4.10). For simplicity of the presentation, we assume that $\theta(z) = (\kappa+\eta)^{-1/2}$ as overall constant factors do not matter (see the remark after (4.10)). We set $\eta = 1/N$, $U = N^{\varepsilon-1}$ and apply Lemma 6.1 to the difference $m^\Delta = m - m_{sc}$. Let $\varrho^\Delta = \varrho - \varrho_{sc}$, where $\varrho(x) = \frac{1}{N}\sum_j \delta(x - \lambda_j)$ is the normalized empirical counting measure of eigenvalues. First we check the conditions of Lemma 6.1. To check that (6.1) holds, set $L = (\log N)^{24+6\alpha}$ and for a fixed

x, let y_x satisfy $Ny_x(\kappa_x+y_x)^{3/2}=L$, so that $x+iy_x\in D^*$. Clearly (6.1) holds for any $y\geq y_x$ with a very high probability by (5.2). In particular, we know that

$$|m(x+iy_x)-m_{sc}(x+iy_x)|\leq\frac{CU}{y_x(\kappa_x+y_x)}. \tag{6.14}$$

Consider $y<y_x$, set $z=x+iy$, $z_x=x+iy_x$ and estimate

$$|m(z)-m_{sc}(z)|\leq|m(z_x)-m_{sc}(z_x)|+\int_y^{y_x}\big|\partial_\eta\big(m(x+i\eta)-m_{sc}(x+i\eta)\big)\big|\mathrm{d}\eta. \tag{6.15}$$

Note that

$$|\partial_\eta m(x+i\eta)|=\Big|\frac{1}{N}\sum_j\partial_\eta G_{jj}(x+i\eta)\Big| \tag{6.16}$$

$$\leq\frac{1}{N}\sum_{jk}|G_{jk}(x+i\eta)|^2=\frac{1}{N\eta}\sum_j\mathfrak{Im}\,G_{jj}(x+i\eta)=\frac{1}{\eta}\mathfrak{Im}\,m(x+i\eta), \tag{6.17}$$

and similarly

$$|\partial_\eta m_{sc}(x+i\eta)|=\Big|\int\frac{\varrho_{sc}(s)}{(s-x-i\eta)^2}\mathrm{d}s\Big|\leq\int\frac{\varrho_{sc}(s)}{|s-x-i\eta|^2}\mathrm{d}s=\frac{1}{\eta}\mathfrak{Im}\,m_{sc}(x+i\eta).$$

Now we use the fact that the functions $y\to y\mathfrak{Im}\,m(x+iy)$ and $y\to y\mathfrak{Im}\,m_{sc}(x+iy)$ are monotone increasing for any $y>0$ since both are Stieltjes transforms of a positive measure. Therefore the integral in (6.15) can be bounded by

$$\int_y^{y_x}\frac{\mathrm{d}\eta}{\eta}\big[\mathfrak{Im}\,m(x+i\eta)+\mathfrak{Im}\,m_{sc}(x+i\eta)\big]\leq y_x\big[\mathfrak{Im}\,m(x+iy_x)+\mathfrak{Im}\,m_{sc}(x+iy_x)\big]\int_y^{y_x}\frac{\mathrm{d}\eta}{\eta^2} \tag{6.18}$$

By the choice of y_x and using that $\mathfrak{Im}\;m_{sc}(z_x)\leq C\sqrt{\kappa_x+y_x}$, we have

$$\mathfrak{Im}\;m_{sc}(z_x)\leq\frac{CU}{y_x(\kappa_x+y_x)}. \tag{6.19}$$

and then $\mathfrak{Im}\;\mathbb{E}m(z_x)$ can be estimated from (6.14). Inserting these estimates into (6.15) and (6.18), and using (6.14), we get

$$|m(z)-m_{sc}(z)|\leq|m(z_x)-m_{sc}(z_x)|+\frac{CU}{y_x(\kappa_x+y_x)}\frac{y_x}{y}\leq\frac{CU}{y(\kappa_x+y)}$$

with a possible larger C in the r.h.s. Thus (6.1) holds for the difference $m^\Delta=m-m_{sc}$.

The application of Lemma 6.1 shows that for $\eta=1/N$

$$\Big|\int f_{E_1,E_2,\eta}(\lambda)\varrho(\lambda)\mathrm{d}\lambda-\int f_{E_1,E_2,\eta}(\lambda)\varrho_{sc}(\lambda)\mathrm{d}\lambda\Big|\leq\frac{CN^{2\varepsilon}}{N\min\{\kappa_{E_1},\kappa_{E_2}\}+1}. \tag{6.20}$$

Recall that $f_{E_1,E_2,\eta}$ the characteristic function of the interval $[E_1,E_2]$, smoothed on scale η at the edges. The additional 1 in the denominator in the r.h.s. of (6.20) comes from the case when κ_{E_1}, κ_{E_2} are very small and the trivial estimate $f\leq1$ with $\int\varrho=\int\varrho_{sc}=1$ gives a better bound than Lemma 6.1.

With the fact $y \to y\mathfrak{Im}\, m(x+iy)$ is monotone increasing for any $y > 0$, (6.19) implies a crude upper bound on the empirical density. Indeed, for any interval $I := [x-\eta, x+\eta]$, with $\eta = 1/N$, we have

$$\mathfrak{n}(x+\eta) - \mathfrak{n}(x-\eta) \leq C\eta\, \mathfrak{Im}\ m\big(x+i\eta\big) \leq Cy_x\, \mathfrak{Im}\ m\big(x+iy_x\big) \leq \frac{CN^{2\varepsilon}}{N\kappa_x + 1}. \tag{6.21}$$

since $\eta = 1/N \leq y_x$ for any x.

Choose arbitrary $E_1, E_2 \in [-3,3]$, then we have

$$\begin{aligned}\Big|\mathfrak{n}(E_1) - \mathfrak{n}(E_2) - \int f_{E_1,E_2,\eta}(\lambda)\varrho(\lambda)\mathrm{d}\lambda\Big| \leq & C\sum_{j=1,2}\big[\mathfrak{n}(E_j+\eta) - \mathfrak{n}(E_j-\eta)\big] \\ \leq & \sum_{j=1,2}\frac{CN^{2\varepsilon}}{N\kappa_{E_j}+1}\end{aligned} \tag{6.22}$$

from (6.21). Since ϱ_{sc} is bounded, we also have

$$\Big|n_{sc}(E_1) - n_{sc}(E_2) - \int f_{E_1,E_2,\eta}(\lambda)\varrho_{sc}(\lambda)\mathrm{d}\lambda\Big| \leq C\eta = C/N. \tag{6.23}$$

Subtracting (6.22) and (6.23) and using (6.20), we obtain that for any $E_1, E_2 \in [-3,3]$

$$\Big|\big[\mathfrak{n}(E_1) - \mathfrak{n}(E_2)\big] - \big[n_{sc}(E_1) - n_{sc}(E_2)\big]\Big| \leq \frac{CN^{2\varepsilon}}{N\min\{\kappa_{E_1}, \kappa_{E_2}\} + 1}$$

with a very high probability, i.e., apart from a set of probability smaller than $C(\varepsilon, K)N^{-K}$ for any K. The estimate (6.13) from Lemma 6.2 on the extreme eigenvalues shows that ϱ is supported in $[-3,3]$ with very high probability, i.e., $\mathfrak{n}(-3) = n_{sc}(-3) = 0$, $\mathfrak{n}(3) = n_{sc}(3) = 1$. Thus we obtain that

$$\Big|\mathfrak{n}(E) - n_{sc}(E)\Big| \leq \frac{CN^{2\varepsilon}}{N\kappa_E + 1} \tag{6.24}$$

holds for any fixed $E \in [-3,3]$ with an overwhelming probability.

We now choose a fine grid of equidistant points $E_j \in [-3,3]$ with $|E_j - E_{j+1}| \leq N^{-1}$, then (6.24) holds simultaneously for every $E = E_j$ with an overwhelming probability. For any $E \in [-3,3]$ we can find an E_j with $|E - E_j| \leq N^{-1}$ and by (6.21) we obtain

$$|\mathfrak{n}(E) - \mathfrak{n}(E_j)| \leq \mathfrak{n}(E_j + 1/N) - \mathfrak{n}(E_j - 1/N) \leq \frac{CN^{2\varepsilon}}{N\kappa_{E_j} + 1}.$$

This guarantees that (6.24) holds simultaneously for all E. Since $\varepsilon > 0$ was arbitrary, this proves Theorem 6.3 for generalized Wigner matrices.

The proof for universal Wigner matrices is very similar, just M replaces N in the estimates, $U = N^\varepsilon M^{-1}$ and instead of $\theta(z) \leq C(\kappa+\eta)^{-1/2}$ one uses $\theta(z) \leq C(\kappa+\eta)^{-1}$ which follows from (4.10). The main technical estimate (6.20) is modified to

$$\Big|\int f_{E_1,E_2,\eta}(\lambda)\varrho(\lambda)\mathrm{d}\lambda - \int f_{E_1,E_2,\eta}(\lambda)\varrho_{sc}(\lambda)\mathrm{d}\lambda\Big| \leq \frac{CN^{2\varepsilon}}{M\big[\min\{\kappa_{E_1}, \kappa_{E_2}\}\big]^2 + 1} \tag{6.25}$$

and the rest of the proof is identical. □

7 Location of eigenvalues

In this section we estimate the mean square deviation of the eigenvalues from their classical location. The main input is Theorem 6.3, the estimate on the counting function. For simplicity, we consider only the case of generalized Wigner matrices. Similar, but weaker results can be obtained along the same lines for universal Wigner matrices.

Theorem 7.1 *Let H be a generalized Wigner matrix with subexponential decay, i.e., assume that (2.1), (2.2), (2.6) and (2.11) hold. Let λ_j denote the eigenvalues of H and γ_j be their classical location, defined by (2.23). Then for any $\varepsilon_0 < 1/7$ and for any $K > 1$ there exists a constant C, depending on K and ε_0, such that*

$$\mathbb{P}\Big\{ \sum_{j=1}^N |\lambda_j - \gamma_j|^2 \le N^{-\varepsilon_0} \Big\} \ge 1 - \frac{C}{N^K}. \tag{7.1}$$

and

$$\sum_{j=1}^N \mathbb{E}|\lambda_j - \gamma_j|^2 \le CN^{-\varepsilon_0}. \tag{7.2}$$

Proof. The proof of (7.2) directly follows from (7.1) by using the estimates on the extreme eigenvalue (6.13) from Lemma 6.2. For the proof of (7.1), we can assume that $\max_j |\lambda_j| \le 2 + N^{-1/7}$ since the complement event has a negligible probability by (6.12) and (6.13) of Lemma 6.2. From Theorem 6.3 we can also assume that

$$|\mathfrak{n}(E) - n_{sc}(E)| \le \frac{CN^\varepsilon}{N\kappa_E} \tag{7.3}$$

holds for every $E \in \mathbb{R}$.

From the definition of γ_j it follows that for $j \le N/2$, i.e., $\gamma_j \le 0$,

$$-2 + C_1\Big(\frac{j}{N}\Big)^{2/3} \le \gamma_j \le -2 + C_2\Big(\frac{j}{N}\Big)^{2/3} \tag{7.4}$$

with some positive constants C_1, C_2.

Choose $\beta = \frac{2}{5} - \varepsilon$. Consider first those j-indices for which $C_0N^{1-3\beta/2} \le j \le N - C_0N^{1-3\beta/2}$ with a sufficiently large constant. We choose C_0 so that (7.4) would imply $-2 + 2N^{-\beta} \le \gamma_j \le 2 - 2N^{-\beta}$. We then claim that

$$\lambda_j \in [-2 + N^{-\beta}, 2 - N^{-\beta}] \qquad \text{for} \quad C_0N^{1-3\beta/2} \le j \le N - C_0N^{1-3\beta/2}. \tag{7.5}$$

We will show that $\lambda_j \ge -2 + N^{-\beta}$, the upper bound is analogous. Suppose that λ_j were smaller than $-2 + N^{-\beta}$, then $\mathfrak{n}(-2 + N^{-\beta}) \ge j$. On the other hand, $n_{sc}(-2 + 2N^{-\beta}) \le j$ and thus

$$n_{sc}(-2 + N^{-\beta}) = n_{sc}(-2 + 2N^{-\beta}) - \int_{-2+N^{-\beta}}^{-2+2N^{-\beta}} \varrho_{sc}(x)\mathrm{d}x \le j - cN^{-3\beta/2}$$

with some positive constant c. Therefore

$$cN^{-3\beta/2} \le \mathfrak{n}(-2 + N^{-\beta}) - n_{sc}(-2 + N^{-\beta}) \le CN^{\beta+\varepsilon-1},$$

where the second inequality follows from (7.3), but this contradicts to the choice $\beta = \frac{2}{5} - \varepsilon$.

Let j satisfy $C_0 N^{1-3\beta/2} \leq j \leq N/2$; the indices $N/2 \leq j \leq N - C_0 N^{1-3\beta/2}$ can be treated analogously. Note that $\lambda_{N/2} \leq CN^{-1+\varepsilon}$ by (7.3). Define $c(j)$ to be index of the γ-point right below λ_j, i.e.,

$$\gamma_{c(j)} \leq \lambda_j \leq \gamma_{c(j)+1}.$$

By (7.5) we see that $-2 + \frac{1}{2}N^{-\beta} \leq \gamma_{c(j)} \leq CN^{-1+\varepsilon}$ and from (7.3) and (7.4) it follows that

$$|c(j) - j| \leq \frac{CN^{\varepsilon}}{2 + \gamma_{c(j)}} \leq CN^{\varepsilon+\beta}. \tag{7.6}$$

By the choice of β we have $\varepsilon + \beta < 1 - \frac{3}{2}\beta$, i.e., (7.6) implies $|c(j) - j| \ll j$. Using now (7.4), we have

$$|c(j) - j| \leq \frac{CN^{\varepsilon}}{\gamma_{c(j)} + 2} \leq \frac{CN^{2/3+\varepsilon}}{c(j)^{2/3}} \leq \frac{CN^{2/3+\varepsilon}}{j^{2/3}}. \tag{7.7}$$

Finally, we can estimate

$$|c(j) - j| = N\Big|\int_{\gamma_{c(j)}}^{\gamma_j} \varrho_{sc}(x)\mathrm{d}x\Big| \geq CN|\gamma_{c(j)} - \gamma_j|(2+\gamma_j)^{1/2} \geq CN|\gamma_{c(j)} - \gamma_j|\Big(\frac{j}{N}\Big)^{1/3},$$

using $|c(j) - j| \ll j$ and hence $(2 + \gamma_j)$ and $(2 + \gamma_{c(j)})$ are comparable. In the last step we also used (7.4). Combining this with (7.7), we have

$$|\gamma_{c(j)} - \gamma_j| \leq C\frac{|c(j) - j|}{N^{2/3}j^{1/3}} \leq \frac{CN^{\varepsilon}}{j}.$$

and the same estimate holds for $|\gamma_{c(j)+1} - \gamma_j|$ and thus

$$|\lambda_j - \gamma_j| \leq \frac{CN^{\varepsilon}}{j}$$

as well. Therefore

$$\sum_{C_0 N^{1-3\beta/2} \leq j \leq N/2} |\lambda_j - \gamma_j|^2 \leq CN^{2\varepsilon - 1 + 3\beta/2} \leq CN^{-2/5+\varepsilon/2} \tag{7.8}$$

by the choice of β and similar estimate holds for the sum over the indices $N/2 \leq j \leq N - C_0 N^{1-3\beta/2}$ as well.

Now we consider the indices $j \leq C_0 N^{1-3\beta/2}$ and $\lambda_j \geq -2 - N^{-\beta}$. By a similar argument that proved (7.5), we can see that there is a constant C_3 such that $\lambda_j \leq -2 + C_3 N^{-\beta}$, otherwise $\mathfrak{n}(-2 + C_3 N^{-\beta}) \leq j$, but $n_{sc}(-2 + C_3 N^{-\beta}) \geq j + cN^{-3\beta/2}$, which would contradict (7.3). It is easy to see that $\gamma_j \leq -2 + CN^{-\beta}$ for all $j \leq C_0 N^{1-3\beta/2}$, therefore in this regime we estimate $|\lambda_j - \gamma_j| \leq CN^{-\beta}$ and thus

$$\sum_{j=1}^{C_0 N^{1-3\beta/2}} |\lambda_j - \gamma_j|^2 \mathbf{1}(\lambda_j \geq -2 - N^{-\beta}) \leq C_0 N^{1-3\beta/2}(CN^{-\beta})^2 \leq CN^{-2/5+7\varepsilon/2}. \tag{7.9}$$

The indices $j \geq N - C_0 N^{1-3\beta/2}$ and $\lambda_j \leq 2 + N^{-\beta}$ can be treated similarly.

Finally we deal with the extreme eigenvalues $\lambda_j \leq -2 - N^{-\beta}$ with index $j \leq C_0 N^{1-3\beta/2}$ and we can assume that $\lambda_j \geq -2 - N^{-1/7}$. For these indices $-2 \leq \gamma_j \leq -2 + CN^{-\beta}$ and we can estimate

$$|\lambda_j - \gamma_j| \leq C|\lambda_j + 2|.$$

For any a with $N^{-\beta} \leq a \leq N^{-1/7}$, we have $n_{sc}(-2-a) = 0$, thus we obtain from (7.3) that

$$\mathfrak{n}(-2-a) \leq \frac{CN^{\varepsilon}}{Na}.$$

Therefore

$$\begin{aligned}\sum_j |\lambda_j - \gamma_j|^2 \mathbf{1}(-2 - N^{-1/7} \leq \lambda_j \leq -2 - N^{-\beta}) \leq& C \sum_j |\lambda_j + 2|^2 \mathbf{1}(-N^{-1/7} \leq \lambda_j + 2 \leq -N^{-\beta}) \\ \leq& C \int_0^{N^{-1/7}} a \cdot \frac{CN^{\varepsilon}}{Na} \, \mathrm{d}a \\ \leq& CN^{-1/7+\varepsilon}. \end{aligned} \tag{7.10}$$

The other extreme eigenvalues, $\lambda_j \geq 2 + N^{-\beta}$, are treated analogously.

Combining (7.8), (7.9) and (7.10) and choosing ε sufficiently small in the definition of β, we proved (7.1) with any $\varepsilon_0 < 1/7$. □

8 Moment Estimates of Error Terms

In this section we prove the second and fourth moment estimates of Lemma 5.2; the general cases will be proved in Section 9.

Definition 8.1 *Define the operator* $\mathbb{IE}_i$ *as*

$$\mathbb{IE}_i \equiv \mathbb{I} - \mathbb{E}_{\mathbf{a}^i}, \tag{8.1}$$

where $\mathbb{I}$ *is identity operator.*

Recall the definition of Z_i, which we rewrite as

$$Z_i = \mathbb{IE}_i Z_{ii}^{(i)}, \qquad Z_{ii}^{(i)} = \sum_{k,l \neq i} \overline{\mathbf{a}_k^i} G_{k\,l}^{(i)} \mathbf{a}_l^i = \mathbf{a}^i \cdot G^{(i)} \mathbf{a}^i. \tag{8.2}$$

We first prove a bound on the Green function $G_{k\,l}^{(i)}$.

Lemma 8.1 *Recall the definition of* X *in* (5.6)*. Let* t *be any fixed positive integer,* $\mathbb{T} = \{k_1, k_2 \ldots k_t\} \in \mathbb{N}^t$*,* $1 \leq k_i \leq N$ *for any* $1 \leq i \leq t$*. Then there exists a constant* C_t*, depending only on* t*, such that for any* $z \in D^*$ *in* (5.1) *in the set* Ω^c (4.33)*, we have*

$$\max_{k,l: l \neq k, \; l,k \notin \mathbb{T}} |G_{lk}^{(\mathbb{T})}(z)| \quad \leq \quad C_t X(z), \tag{8.3}$$

$$\max_{k: k \notin \mathbb{T}} |G_{kk}^{(\mathbb{T})}(z) - m_{sc}(z)| \quad \leq \quad C_t X(z) \theta(z) \tag{8.4}$$

and for some constant c, C independent of t,

$$c \le \min_{k:k\notin\mathbb{T}} |G^{(\mathbb{T})}_{kk}(z)| \le \max_{k:k\notin\mathbb{T}} |G^{(\mathbb{T})}_{kk}(z)| \le C, \tag{8.5}$$

for sufficiently large N.

Proof Consider first the case $t=0$. Let Y denote the event inside the probability in the equation (4.3). The proof of Theorem 4.1 yields that $\Omega^c \subset Y^c$. It is clear that (8.4) holds in the event Y^c and this proves (8.4) in Ω^c in the case $t=0$. Similarly, in the case of $t=0$, we can prove (8.3) using the event in the equation (4.4). By definition of the domain D^*, the right side of (8.4) is $o(1)$ and this proves (8.5) in the case $t=0$.

For the case $t=1$ and $i_1=i$, using (4.15) and (4.16), we obtain that

$$\begin{aligned} |G^{(i)}_{lk}| &\le |G_{lk}| + |G_{li}G_{ik}||G_{ii}|^{-1}, \\ |G^{(i)}_{kk} - m_{sc}| &\le |G_{kk} - m_{sc}| + |G_{ki}G_{ik}||G_{ii}|^{-1}. \end{aligned} \tag{8.6}$$

Since $X^2 \ll X$ in D^*, (8.4) and (8.3) in the case $t=1$ follows from (8.6) and the case $t=0$. Repeating this process, we prove (8.4) and (8.3) for any $t>1$ by induction on t. □

Now we return to the second and fourth moment estimates of Lemma 5.2.

8.1 Proof of Lemma 5.2 for $p=2$.

Now we prove the special case of Lemma 5.2 for $p=2$. The second moment of $\sum_{i=1}^N Z_i$ is given by

$$\frac{1}{N^2}\mathbb{E}\left|\sum_{i=1}^N Z_i\right|^2 = \frac{1}{N^2}\mathbb{E}\sum_{\alpha\neq\beta}\overline{Z_\alpha}Z_\beta + \frac{1}{N^2}\mathbb{E}\sum_\alpha |Z_\alpha|^2. \tag{8.7}$$

We start with estimating the first term of (8.7) for $\alpha=1$ and $\beta=2$. The basic idea is to rewrite $G^{(1)}_{k\,l}$ as

$$G^{(1)}_{k\,l} = P^{(1),\emptyset}_{k\,l} + P^{(1),(2)}_{k\,l}, \qquad k,l\neq 1, \tag{8.8}$$

with $P^{(1),(2)}_{k\,l}$ independent of $\mathbf{a}^1$, $\mathbf{a}^2$ and $P^{(1),\emptyset}_{k\,l}$ independent of $\mathbf{a}^1$. The P's have two upper indices. The first one refers to the fact that it comes from the $H^{(1)}$ minor (i.e. follows the upper index of $G^{(1)}$) and the second one indicates the additional independence.

To construct this decomposition for $k,l\notin\{1,2\}$, by (4.15) or (4.16) we can rewrite $G^{(1)}_{k\,l}$ as

$$G^{(1)}_{k\,l} = G^{(12)}_{k\,l} + \frac{G^{(1)}_{k\,2}G^{(1)}_{2\,l}}{G^{(1)}_{2\,2}}, \qquad k,l\notin\{1,2\}. \tag{8.9}$$

The first term on the r.h.s is independent of $\mathbf{a}^2$. With Lemma 8.1, we have that the bound

$$\left|\frac{G^{(1)}_{k\,2}G^{(1)}_{2\,l}}{G^{(1)}_{2\,2}}\right| \le CX^2 \tag{8.10}$$

holds with a very high probability.

Next we define $P^{(1)}$ for $(k,l\neq 1)$.

1. If $k, l \neq 2$,

$$P^{(1),(2)}_{k\,l} = G^{(12)}_{k\,l}, \;\; P^{(1),\emptyset}_{k\,l} = \frac{G^{(1)}_{k\,2}G^{(1)}_{2\,l}}{G^{(1)}_{2\,2}} = G^{(1)}_{k\,l} - G^{(12)}_{k\,l}. \tag{8.11}$$

2. If $k = 2$ or $l = 2$,

$$P^{(1),(2)}_{k\,l} = 0, \;\; P^{(1),\emptyset}_{k\,l} = G^{(1)}_{k\,l}. \tag{8.12}$$

Hence (8.8) holds and $P^{(1),(2)}_{k\,l}$ is independent of $\mathbf{a}^2$.

With this convention, we have the following expansion of Z_1

$$Z_1 = \mathbb{IE}_1\mathbf{a}^1 \cdot P^{(1),(2)}\mathbf{a}^1 + \mathbb{IE}_1\mathbf{a}^1 \cdot P^{(1),\emptyset}\mathbf{a}^1. \tag{8.13}$$

Lemma 8.2 *For $N^{-1} \leq \eta \leq 10$ and fixed $p \in \mathbb{N}$, we have the following estimates*

$$\mathbb{E}\left|\mathbf{a}^1 \cdot P^{(1),\emptyset}\mathbf{a}^1\right|^p \leq C_p\left((\log N)^{3+2\alpha}\right)^p X^{2p}, \tag{8.14}$$

$$\mathbb{E}\left|\mathbf{a}^1 \cdot P^{(1),(2)}\mathbf{a}^1\right|^p \leq C_p\left((\log N)^{3+2\alpha}\right)^p X^{p}. \tag{8.15}$$

Since $X^2 \leq X$ in D^*, this lemma also implies that

$$\mathbb{E}\,|Z_i|^p \leq C_k\left((\log N)^{3+2\alpha}\right)^p X^p, \qquad 1 \leq i \leq N. \tag{8.16}$$

Proof. First we rewrite $\mathbf{a}^1 \cdot P^{(1),\emptyset}\mathbf{a}^1$ as follows

$$\mathbf{a}^1 \cdot P^{(1),\emptyset}\mathbf{a}^1 = \sum_{k,l\neq 2}\overline{\mathbf{a}^1_k}\left(\frac{G^{(1)}_{k\,2}G^{(1)}_{2\,l}}{G^{(1)}_{2\,2}}\right)\mathbf{a}^1_l + \sum_{k\neq 2}\overline{\mathbf{a}^1_k}G^{(1)}_{k\,2}\mathbf{a}^1_2 + \sum_{l\neq 2}\overline{\mathbf{a}^1_2}G^{(1)}_{2\,l}\mathbf{a}^1_l + \overline{\mathbf{a}^1_2}G^{(1)}_{2\,2}\mathbf{a}^1_2. \tag{8.17}$$

By the large deviation estimate (4.19), we have

$$\mathbb{P}\left(\left|\sum_{k,l\neq 2}\overline{\mathbf{a}^1_k}\left(\frac{G^{(1)}_{k\,2}G^{(1)}_{2\,l}}{G^{(1)}_{2\,2}}\right)\mathbf{a}^1_l\right| \geq C(\log N)^{3+2\alpha}X^2\right) \leq N^{-c\log\log N}. \tag{8.18}$$

Similarly, from (4.17), using a_i as $\mathbf{a}^1_3$, $\mathbf{a}^1_4, \ldots, \mathbf{a}^1_N$ and keeping $\mathbf{a}^1_2$ fixed, we have

$$\mathbb{P}\left(\left|\sum_{k\neq 2}\overline{\mathbf{a}^1_k}G^{(1)}_{k\,2}\mathbf{a}^1_2\right| \geq C(\log N)^{3/2+\alpha}X|\mathbf{a}^1_2|\right) \leq N^{-c\log\log N}. \tag{8.19}$$

By (4.35), $\|\mathbf{a}^1\|_\infty \leq (\log N)^{2\alpha}M^{-1/2}$ holds with a very high probability. We can thus replace $|\mathbf{a}^1_2|$ by $(\log N)^{2\alpha}M^{-1/2}$ in (8.19). The third term in (8.17) can be estimated in the same way, and the last term can be bounded by $(\log N)^{4\alpha}\frac{1}{M}$ with very high probability.

Since $\eta \leq 10$, by the definition of X in (5.6) we have

$$X^2 \geq C(\log N)^2/M. \tag{8.20}$$

Thus

$$(\log N)^{3/2+3\alpha} \frac{X}{\sqrt{M}} + (\log N)^{4\alpha} \frac{1}{M} \le C (\log N)^{3+2\alpha} X^2,$$

and we have proved that

$$\mathbb{P}\left(\left| \mathbf{a}^1 \cdot P^{(1),\emptyset} \mathbf{a}^1 \right| \le C(\log N)^{3+2\alpha} X^2 \right) \ge 1 - N^{-c \log\log N}. \tag{8.21}$$

This inequality implies the desired inequality (8.14) except for the contribution from the exceptional set where (8.21) fails. Since all Green functions are bounded by $\eta^{-1} \le N$, the contribution from the exceptional set is negligible and this proves (8.14). Finally, a similar proof yields (8.15). □

Exchange the index 1 and 2, we can define $P^{(2),(1)}$ and $P^{(2),\emptyset}$ and expand Z_2 as

$$Z_2 = \mathbb{IE}_2 \mathbf{a}^2 \cdot P^{(2),(1)} \mathbf{a}^2 + \mathbb{IE}_2 \mathbf{a}^2 \cdot P^{(2),\emptyset} \mathbf{a}^2. \tag{8.22}$$

Here $P^{(2),(1)}_{k\,l}$ is independent of $\mathbf{a}^2$ and $\mathbf{a}^1$; $P^{(2),\emptyset}_{k\,l}$ is independent of $\mathbf{a}^2$. Combining (8.22) with (8.13), we have

$$\mathbb{E} \overline{Z}_1 Z_2 = \mathbb{E}\left[\left(\mathbb{IE}_1 \left\{ \overline{\mathbf{a}^1 \cdot P^{(1),(2)} \mathbf{a}^1 + \mathbf{a}^1 \cdot P^{(1),\emptyset} \mathbf{a}^1} \right\} \right) \left(\mathbb{IE}_2 \left\{ \mathbf{a}^2 \cdot P^{(2),(1)} \mathbf{a}^2 + \mathbf{a}^2 \cdot P^{(2),\emptyset} \mathbf{a}^2 \right\} \right) \right]. \tag{8.23}$$

The only non-vanishing term on the right-hand side is

$$\mathbb{E}\left(\mathbb{IE}_1 \overline{\mathbf{a}^1 \cdot P^{(1),\emptyset} \mathbf{a}^1} \right) \left(\mathbb{IE}_2 \mathbf{a}^2 \cdot P^{(2),\emptyset} \mathbf{a}^2 \right). \tag{8.24}$$

By the Cauchy-Schwarz inequality and Lemma 8.2, we obtain

$$|\mathbb{E} \overline{Z}_1 Z_2| \le C \left((\log N)^{3+2\alpha} \right)^2 X^4. \tag{8.25}$$

Similarly, Lemma 8.2 and (8.20) imply that

$$\mathbb{E} |Z_1|^2 \le C \left((\log N)^{3+2\alpha} \right)^2 X^2 \le CM \left((\log N)^{3+2\alpha} \right)^2 X^4.$$

Since the indices 1 and 2 can be replaced by $\alpha \ne \beta$, together with (8.7) we have thus proved Lemma 5.2 for $p = 2$.

8.2 Proof of Lemma 5.2 for $p = 4$

Now we prove the special case of Lemma 5.2 for $p = 4$:

$$\begin{aligned} N^{-4} \mathbb{E} \left| \sum_{i=1}^N Z_i \right|^4 \le\; & CN^{-4} \sum_{1 \le \alpha < \beta < \chi < \gamma \le N} \left| \mathbb{E}\, \overline{Z}_\alpha \overline{Z}_\beta Z_\chi Z_\gamma \right| \\ & + CN^{-4} \sum_{1 \le \alpha < \beta < \chi \le N} \left| \mathbb{E} |Z_\alpha|^2 \overline{Z}_\beta Z_\chi \right| + \dots \\ & + CN^{-4} \sum_{1 \le \alpha < \beta \le N} \left(\mathbb{E} |Z_\alpha|^2 |Z_\beta|^2 + \left| \mathbb{E} |Z_\alpha|^2 \overline{Z}_\alpha Z_\beta \right| \right) + \dots \\ & + CN^{-4} \sum_{1 \le \alpha \le N} \mathbb{E} |Z_\alpha|^4. \end{aligned} \tag{8.26}$$

Here ... means the permutation of the ordered indices and the complex conjugate operators. We are going to compute the first two terms in the r.h.s of (8.26). The other two terms can be treated analogously. By the permutation symmetry of the indices, we can assume that $\alpha = 1$, $\beta = 2$, $\chi = 3$ and $\gamma = 4$. As in the estimate for the second moment, the key idea is to decompose $Z_{11}^{(1)}$ in a suitable way:

Lemma 8.3 *There exist two decompositions of* $Z_{11}^{(1)}$

$$Z_{11}^{(1)} = \sum_{\mathbb{T}\subset\{2,3\}} \mathbf{a}^1 \cdot Q^{(1),(\mathbb{T})}\mathbf{a}^1, \quad Z_{11}^{(1)} = \sum_{\mathbb{T}\subset\{2,3,4\}} \mathbf{a}^1 \cdot R^{(1),(\mathbb{T})}\mathbf{a}^1, \tag{8.27}$$

such that $Q^{(1),(\mathbb{T})}$ *and* $R^{(1),(\mathbb{T})}$ *are independent of the rows in* $\mathbb{T}\cup\{1\}$*, i.e.,*

$$\frac{\partial\left(\mathbf{a}^1 \cdot Q^{(1),(\mathbb{T})}\mathbf{a}^1\right)}{\partial \mathbf{a}^i_j} = \frac{\partial\left(\mathbf{a}^1 \cdot Q^{(1),(\mathbb{T})}\mathbf{a}^1\right)}{\partial \overline{\mathbf{a}^i_j}} = 0, \quad i\in\mathbb{T}\subset\{2,3\},\ 1\le j\le N. \tag{8.28}$$

and

$$\frac{\partial\left(\mathbf{a}^1 \cdot R^{(1),(\mathbb{T})}\mathbf{a}^1\right)}{\partial \mathbf{a}^i_j} = \frac{\partial\left(\mathbf{a}^1 \cdot R^{(1),(\mathbb{T})}\mathbf{a}^1\right)}{\partial \overline{\mathbf{a}^i_j}} = 0, \quad i\in\mathbb{T}\subset\{2,3,4\},\ 1\le j\le N. \tag{8.29}$$

Furthermore, the decompositions can be chosen in such a way that for all $N^{-1}\le\eta\le 10$ *the following estimates hold:*

$$\mathbb{E}\left|\mathbf{a}^1 \cdot Q^{(1),(\mathbb{T})}\mathbf{a}^1\right|^p \le C_p\left((\log N)^{3+2\alpha}\right)^p (X^{3-|\mathbb{T}|})^p, \quad p\in\mathbb{N} \tag{8.30}$$

and

$$\mathbb{E}\left|\mathbf{a}^1 \cdot R^{(1),(\mathbb{T})}\mathbf{a}^1\right|^p \le C_p\left((\log N)^{3+2\alpha}\right)^p (X^{4-|\mathbb{T}|})^p, \quad p\in\mathbb{N}. \tag{8.31}$$

We postpone the proof of this lemma and first finish the proof of Lemma 5.2 in the case of $p = 4$. It is clear that Lemma 8.3 holds for different index combinations. E.g. $Z_{22}^{(2)}$ can be decomposed as

$$Z_{22}^{(2)} = \sum_{\mathbb{T}\subset\{1,3,4\}} \mathbf{a}^2 \cdot R^{(2),(\mathbb{T})}\mathbf{a}^2 \tag{8.32}$$

and $R^{(2)}$'s have the same properties (except for the exchange of 1 and 2) as $R^{(1)}$ in (8.29) and (8.31) . By this property, we can estimate the first term on the r.h.s. of (8.26) by

$$\begin{aligned}
&\mathbb{E}\left(\mathbb{IE}_1\overline{Z_{11}^{(1)}}\right)\left(\mathbb{IE}_2\overline{Z_{22}^{(2)}}\right)\left(\mathbb{IE}_3 Z_{33}^{(3)}\right)\left(\mathbb{IE}_4 Z_{44}^{(4)}\right) \\
\le\quad &\mathbb{E}\overline{\left[\mathbb{IE}_1\sum_{\mathbb{T}_1\subset\{2,3,4\}}\mathbf{a}^1\cdot R^{(1),(\mathbb{T}_1)}\mathbf{a}^1\right]}\times\overline{\left[\mathbb{IE}_2\sum_{\mathbb{T}_2\subset\{1,3,4\}}\mathbf{a}^2\cdot R^{(2),(\mathbb{T}_2)}\mathbf{a}^2\right]}\left[\cdots R^{(3),(\mathbb{T}_3)}\cdots\right]\left[\cdots R^{(4),(\mathbb{T}_4)}\cdots\right].
\end{aligned} \tag{8.33}$$

Consider a term consisting of products of factors with $\cap_{j=1,2,3,4}(\mathbb{T}_j\cup\{j\})\ne\emptyset$. Then there is an element $\ell\in\{1,2,3,4\}$ in the common intersection so that integration w.r.t. the row $\mathbf{a}^\ell$ vanishes. Hence the nonvanishing terms consist of products of term with $\cap_{j=1,2,3,4}(\mathbb{T}_j\cup\{j\})=\emptyset$, i.e., $\cup_{j=1,2,3,4}(\mathbb{T}_j\cup\{j\})^c=\{1,2,3,4\}$. Here the notation ^{c} means the complement in $\{1,2,3,4\}$. Thus we have

$$\sum_{j=1}^4(4-|\mathbb{T}_j|-1)\ge 4 \Longrightarrow \sum_{j=1}^4 4-|\mathbb{T}_j|\ge 8.$$

Using (8.31) and Schwarz inequality, we have thus proved that

$$\left|\mathbb{E}\,\left(\mathbb{IE}_1\overline{Z^{(1)}_{11}}\right)\,\left(\mathbb{IE}_2\overline{Z^{(2)}_{22}}\right)\,\left(\mathbb{IE}_3 Z^{(3)}_{33}\right)\,\left(\mathbb{IE}_4 Z^{(4)}_{44}\right)\right|\le C\left((\log N)^{3+2\alpha}\right)^4 X^8.$$

We now estimate the second term in r.h.s of (8.26).

$$\begin{aligned}\mathbb{E}\,\left|\mathbb{IE}_1 Z^{(1)}_{11}\right|^2\,\left(\mathbb{IE}_2\overline{Z^{(2)}_{22}}\right)\,\left(\mathbb{IE}_3 Z^{(3)}_{33}\right)=\mathbb{E}&\left(\overline{\mathbb{IE}_1\sum_{\mathbb{T}_0\subset\{2,3\}}\mathbf{a}^1\cdot Q^{(1),(\mathbb{T}_0)}\mathbf{a}^1}\right)\,\left(\mathbb{IE}_1\sum_{\mathbb{T}_1\subset\{2,3\}}\mathbf{a}^1\cdot Q^{(1),(\mathbb{T}_1)}\mathbf{a}^1\right)\\&\times\left[\overline{\mathbb{IE}_2\sum_{\mathbb{T}_2\subset\{1,3\}}\mathbf{a}^2\cdot Q^{(2),(\mathbb{T}_2)}\mathbf{a}^2}\right]\times\left[\mathbb{IE}_3\sum_{\mathbb{T}_3\subset\{1,2\}}\mathbf{a}^3\cdot Q^{(3),(\mathbb{T}_3)}\mathbf{a}^3\right]\end{aligned}\tag{8.34}$$

Consider a term consisting of products of factors with $[\cap_{j=2,3}(\mathbb{T}_j\cup\{j\}]\cap\mathbb{T}_0\cap\mathbb{T}_1\neq\emptyset$. Then there is an element $\ell\in\{2,3\}$ in the common intersection and the integration w.r.t. the row $\mathbf{a}^\ell$ vanishes. Thus the nonvanishing terms consist of products of term with $[\cap_{j=2,3}(\mathbb{T}_j\cup\{j\}]\cap\mathbb{T}_0\cap\mathbb{T}_1=\emptyset$. In particular, $\{2,3\}\subset\cup_{j=2,3}(\mathbb{T}_j\cup\{j\})^c\cup[\{2,3\}\setminus\mathbb{T}_0]\cup[\{2,3\}\setminus\mathbb{T}_1]$. Here the notation ^{c} means the complement in $\{1,2,3\}$. Thus we have

$$\sum_{j=0}^{3}(2-|\mathbb{T}_j|)\ge 2\Longrightarrow\sum_{j=0}^{3}3-|\mathbb{T}_j|\ge 6.$$

Using (8.30), (8.20) and a Schwarz inequality, we have

$$N^{-1}\left|\mathbb{E}\,\left|\mathbb{IE}_1 Z^{(1)}_{11}\right|^2\,\left(\mathbb{IE}_2\overline{Z^{(2)}_{22}}\right)\,\left(\mathbb{IE}_3 Z^{(3)}_{33}\right)\right|\le\frac{C}{N}\left((\log N)^{3+2\alpha}\right)^4 X^6\ll C\left((\log N)^{3+2\alpha}\right)^4 X^8.$$

For the other terms in (8.26), we can just use Schwarz inequality and (8.16). We have thus proved the Lemma 5.2 for $p=4$.

We now prove Lemma 8.3. First we prove the properties of Q's. Notice that the decomposition with Q's in (8.27) removes the dependence on rows $2,3$. The starting point is an expansion of $G^{(1)}_{k\,l}$

$$G^{(1)}_{k\,l}=\sum_{\mathbb{T}\subset\{2,3\}}Q^{(1),(\mathbb{T})}_{k\,l}=Q^{(1),\emptyset}_{k\,l}+Q^{(1),(2)}_{k\,l}+Q^{(1),(3)}_{k\,l}+Q^{(1),(2,3)}_{k\,l},\tag{8.35}$$

where $Q^{(1),\mathbb{T}}_{k\,l}$ is independent of the rows and columns in $\mathbb{T}\cup\{1\}$. Using the notation $(1\,\mathbb{U})$ for $(\{1\}\cup\mathbb{U})$, one can check that a solution for Q is given by

$$Q^{(1),(\mathbb{T})}_{k\,l}=\sum_{\mathbb{U}:\mathbb{T}\subset\mathbb{U}\subset\{2,3\}\setminus\{k,l\}}(-1)^{|\mathbb{U}|-|\mathbb{T}|}G^{(1\mathbb{U})}_{k\,l}.\tag{8.36}$$

Thus $Q^{(1),(\mathbb{T})}_{k\,l}=0$ if k or $l\in\mathbb{T}$. By definition of $Z^{(1)}_{11}$ (8.2) and (8.35), we have that the Q's satisfy (8.27). For any fixed $\mathbb{T}$, $Q^{(1),\mathbb{T}}_{k\,l}$ is independent of the rows (column) in $\mathbb{T}\cup\{1\}$. Thus we proved (8.28).

In order to prove (8.30), we give another representation of the Q's. We begin by removing the dependence of the (kl) matrix element of the Green function on the index 3 for $k,l>3$. By (4.15) or (4.16), we can rewrite the first term of r.h.s of (8.9) as

$$G^{(12)}_{k\,l}=G^{(123)}_{k\,l}+\frac{G^{(12)}_{k\,3}G^{(12)}_{3\,l}}{G^{(12)}_{3\,3}},\qquad k,l\notin\{1,2,3\}.\tag{8.37}$$

This removes the dependence of $G^{(12)}_{k\,l}$ on the index 3 with the last term as the error term. For the last term on r.h.s of (8.9), using (4.15) and (4.16) again, we have

$$G^{(1)}_{k\,2} = G^{(13)}_{k\,2} + \frac{G^{(1)}_{k\,3}G^{(1)}_{3\,2}}{G^{(1)}_{3\,3}}, \qquad G^{(1)}_{2\,l} = G^{(13)}_{2\,l} + \frac{G^{(1)}_{2\,3}G^{(1)}_{3\,l}}{G^{(1)}_{3\,3}}, \qquad G^{(1)}_{2\,2} = G^{(13)}_{2\,2} + \frac{G^{(1)}_{2\,3}G^{(1)}_{3\,2}}{G^{(1)}_{3\,3}}. \tag{8.38}$$

The last equality implies

$$\frac{1}{G^{(1)}_{2\,2}} = \frac{1}{G^{(13)}_{2\,2}} - \frac{G^{(1)}_{2\,3}G^{(1)}_{3\,2}}{G^{(1)}_{3\,3}G^{(1)}_{2\,2}G^{(13)}_{2\,2}}. \tag{8.39}$$

This removes the dependence on the index 3 of both the Green functions and their inverse in the last term in (8.9). Inserting (8.37)–(8.39) into (8.9), we obtain that if $k, l \notin \{1,2,3\}$

$$G^{(1)}_{k\,l} = G^{(123)}_{k\,l} + \frac{G^{(12)}_{k\,3}G^{(12)}_{3\,l}}{G^{(12)}_{3\,3}} + \left(G^{(13)}_{k\,2} + \frac{G^{(1)}_{k\,3}G^{(1)}_{3\,2}}{G^{(1)}_{3\,3}}\right)\left(G^{(13)}_{2\,l} + \frac{G^{(1)}_{2\,3}G^{(1)}_{3\,l}}{G^{(1)}_{3\,3}}\right)\left(\frac{1}{G^{(13)}_{2\,2}} - \frac{G^{(1)}_{2\,3}G^{(1)}_{3\,2}}{G^{(1)}_{3\,3}G^{(1)}_{2\,2}G^{(13)}_{2\,2}}\right). \tag{8.40}$$

So for $k, l \notin \{1,2,3\}$, we define $Q^{(1),\mathbb{T}}_{kl}$ as follows

$$Q^{(1),(2,3)}_{kl} = G^{(123)}_{k\,l}, \qquad Q^{(1),(2)}_{kl} = \frac{G^{(12)}_{k\,3}G^{(12)}_{3\,l}}{G^{(12)}_{3\,3}}, \qquad Q^{(1),(3)}_{kl} = \frac{G^{(13)}_{k\,2}G^{(13)}_{2\,l}}{G^{(13)}_{2\,2}},$$

$$\begin{aligned} Q^{(1),\emptyset}_{kl} = \quad & \frac{G^{(1)}_{k\,3}G^{(1)}_{3\,2}G^{(13)}_{2\,l}}{G^{(1)}_{3\,3}G^{(13)}_{2\,2}} + \frac{G^{(13)}_{k\,2}G^{(1)}_{2\,3}G^{(1)}_{3\,l}}{G^{(1)}_{3\,3}G^{(13)}_{2\,2}} + \frac{G^{(1)}_{k\,3}G^{(1)}_{3\,2}G^{(1)}_{2\,3}G^{(1)}_{3\,l}}{G^{(1)}_{3\,3}G^{(1)}_{3\,3}G^{(13)}_{2\,2}} - \frac{G^{(13)}_{k\,2}G^{(1)}_{2\,3}G^{(1)}_{3\,2}G^{(13)}_{2\,l}}{G^{(1)}_{3\,3}G^{(1)}_{2\,2}G^{(13)}_{2\,2}} \\ & - \frac{G^{(1)}_{k\,3}G^{(1)}_{3\,2}G^{(1)}_{2\,3}G^{(1)}_{3\,2}G^{(13)}_{2\,l}}{G^{(1)}_{3\,3}G^{(1)}_{2\,2}G^{(13)}_{2\,2}G^{(1)}_{3\,3}} - \frac{G^{(13)}_{k\,2}G^{(1)}_{2\,3}G^{(1)}_{3\,2}G^{(1)}_{2\,3}G^{(1)}_{3\,l}}{G^{(1)}_{3\,3}G^{(1)}_{2\,2}G^{(13)}_{2\,2}G^{(1)}_{3\,3}} - \frac{G^{(1)}_{k\,3}G^{(1)}_{3\,2}G^{(1)}_{2\,3}G^{(1)}_{3\,2}G^{(1)}_{2\,3}G^{(1)}_{3\,l}}{G^{(1)}_{3\,3}G^{(1)}_{2\,2}G^{(13)}_{2\,2}G^{(1)}_{3\,3}G^{(1)}_{3\,3}}. \end{aligned} \tag{8.41}$$

One can see that in this case, $k, l \notin \{1,2,3\}$, (8.35) holds and $Q^{(1),(\mathbb{T})}_{k\,l}$'s are independent of the rows (column) in $\mathbb{T} \cup \{1\}$. For $k = 2,3$ or $l = 2,3$ the previous formulas for Q do not make sense. But in this case, we do not need to decompose $G^{(1)}$ in such fine details and we will use the simple decomposition

$$\begin{aligned} G^{(1)}_{2\,l} &= G^{(13)}_{2\,l} + \frac{G^{(1)}_{2\,3}G^{(1)}_{3\,l}}{G^{(1)}_{3\,3}}, \qquad l \neq 3 \ \text{ and } \ G^{(1)}_{2\,l} = G^{(1)}_{2\,3}, \qquad l = 3, \\ G^{(1)}_{3\,l} &= G^{(12)}_{3\,l} + \frac{G^{(1)}_{3\,2}G^{(1)}_{2\,l}}{G^{(1)}_{2\,2}}, \qquad l \neq 2 \ \text{ and } \ G^{(1)}_{3\,l} = G^{(1)}_{3\,2}, \qquad l = 2. \end{aligned}$$

More precisely, we define $Q^{(1),(\mathbb{T})}$ by

1. For $k = 2$ and $l \neq 3$, $Q^{(1),(2,3)}_{k\,l} = Q^{(1),(2)}_{k\,l} = 0$, $Q^{(1),(3)}_{k\,l} = G^{(13)}_{k\,l}$ and $Q^{(1),\emptyset}_{k\,l} = \frac{G^{(1)}_{k\,3}G^{(1)}_{3\,l}}{G^{(1)}_{3\,3}}$.

2. For $k = 2$ and $l = 3$, $Q^{(1),(2,3)}_{k\,l} = Q^{(1),(2)}_{k\,l} = Q^{(1),(3)}_{k\,l} = 0$ and $Q^{(1),\emptyset}_{k\,l} = G^{(1)}_{k\,l}$.

3. For $k = 3$ and $l \neq 2$, $Q^{(1),(3,2)}_{k\,l} = Q^{(1),(3)}_{k\,l} = 0$, $Q^{(1),(2)}_{k\,l} = G^{(12)}_{k\,l}$ and $Q^{(1),\emptyset}_{k\,l} = \frac{G^{(1)}_{k\,2}G^{(1)}_{2\,l}}{G^{(1)}_{2\,2}}$.

4. For $k=3$ and $l=2$, $Q^{(1),(3,2)}_{k\,l} = Q^{(1),(3)}_{k\,l} = Q^{(1),(2)}_{k\,l} = 0$ and $Q^{(1),\emptyset}_{k\,l} = G^{(1)}_{k\,l}$.

Similarly, we can define $Q^{(1),(\mathbb{T})}$ for the cases $l=2$ or $l=3$. We now list the properties of $Q^{(1),(\mathbb{T})}_{kl}$ for $k,l>1$ and $\mathbb{T}\subset\{2,3\}$:

1. $Q^{(1),(\mathbb{T})}_{k\,l}$'s are independent of rows (column) in $\mathbb{T}\cup\{1\}$ and (8.35) holds.
2.
$$Q^{(1),(\mathbb{T})}_{k\,l} = 0 \quad \text{if } k \text{ or } l\in\mathbb{T}. \tag{8.42}$$
3. If $k=l$ and $\mathbb{T}\cup\{k\}=\{2,3\}$, then
$$Q^{(1),(\mathbb{T})}_{k\,l} = G^{(123)}_{k\,l}. \tag{8.43}$$
For all other cases, $Q^{(1),(\mathbb{T})}_{k\,l}$ is a finite sum of terms of the form:
$$\frac{G_oG_o\cdots G_o}{G_dG_d\cdots G_d} \tag{8.44}$$
where each G_o (G_d resp.) represents some off-diagonal (diagonal resp.) matrix element of $G^{(\mathbb{U})}$ with $\mathbb{U}$ some finite set. Furthermore, for $k\neq l$ or $\mathbb{T}\cup\{k\}\neq\{2,3\}$, the number of the off-diagonal elements in the numerator of (8.44) is strictly bigger than $|\{2,3\}\backslash(\mathbb{T}\cup\{k,l\})|$. Using Lemma 8.1, in the set Ω^c we have
$$|Q^{(1),(\mathbb{T})}_{k\,l}| \leq C\left(X^{|\{2,3\}\backslash(\mathbb{T}\cup\{k,l\})|+1} + \mathbf{1}(\mathbb{T}\cup\{k\}=\{2,3\}, k=l)\right). \tag{8.45}$$

Since the probability of the exceptional set Ω is extremely small, a simple argument which we repeated many times shows that it can be neglected in the estimate of the expectation in (8.30). Hence (8.30) follows from (8.45).

The proof of (8.30) shows clearly the approach to remove an element one by one from the Green function. Define $R^{(1),(\mathbb{T})}_{k\,l}$ as follows (like Q's in (8.36))

$$R^{(1),(\mathbb{T})}_{k\,l} \equiv \sum_{\mathbb{U}:\mathbb{T}\subset\mathbb{U}\subset\{2,3,4\}\backslash\{k,l\}} (-1)^{|\mathbb{U}|-|\mathbb{T}|} G^{(1\mathbb{U})}_{k\,l}. \tag{8.46}$$

Using the same method we used for Q's, one can prove the properties of R's in Lemma 8.3. The details will be omitted since we will prove the general cases in the next section.

9 General case

The first step to prove the general cases of Lemma 5.2 is to extend the decomposition (8.27). For any fixed i, $1\leq i\leq N$, and a fixed set $\mathbb{S}=\{i_1,i_2,\ldots,i_s\}$ such that $i\notin\mathbb{S}$, $1\leq i_j\leq N$, our goal is to decompose $Z^{(i)}_{ii}$ so that the following lemma holds:

Lemma 9.1 *For $i\notin\mathbb{S}$, $\mathbb{T}\subset\mathbb{S}$ and $\eta\geq 1/N$, there is a decomposition of*

$$Z^{(i)}_{ii} = \sum_{\mathbb{T}\subset\mathbb{S}} \mathcal{Z}^{(i),\mathbb{S},(\mathbb{T})}, \qquad \mathcal{Z}^{(i),\mathbb{S},(\mathbb{T})} \equiv \sum_{k,l} \overline{\mathbf{a}}^i_k \mathcal{G}^{(i),\mathbb{S},(\mathbb{T})}_{k\,l} \mathbf{a}^i_l. \tag{9.1}$$

such that
(1) $\mathcal{G}^{(i),\mathbb{S},(\mathbb{T})}_{k\,l}$ *is independent of the rows or columns of* H *in* $\{i\}\cup\mathbb{T}$*, i.e.,*

$$\frac{\partial \mathcal{G}^{(i),\mathbb{S},(\mathbb{T})}_{k\,l}}{\partial \mathbf{a}^a_b}=0,\ \ \frac{\partial \mathcal{G}^{(i),\mathbb{S},(\mathbb{T})}_{k\,l}}{\partial \overline{\mathbf{a}^a_b}}=0,\ \ a\in\{i\}\cup\mathbb{T},\ \ 1\le b\le N. \tag{9.2}$$

(2) For any positive integer k*,*

$$\mathbb{E}\left|\mathcal{Z}^{(i),\mathbb{S},(\mathbb{T})}\right|^k\le C_{k,s}\left((\log N)^{3+2\alpha}\right)^k(X^{s-t+1})^k,\quad s=|\mathbb{S}|,\ \ t=|\mathbb{T}|. \tag{9.3}$$

In the applications, $\mathbb{S}$ will be the set of indices, the dependencies of which we wish to isolate in $Z^{(i)}_{ii}$. For example, for the case $i=1$ and $\mathbb{S}=\{2\}$ or $\mathbb{S}=\{2,3\}$, respectively, if we define

$$\mathcal{Z}^{(1),\{2\},(\mathbb{T})}=\mathbf{a}^1\cdot P^{(1),(\mathbb{T})}\mathbf{a}^1,\qquad \mathcal{Z}^{(1),\{2,3\},(\mathbb{T})}\equiv\mathbf{a}^1\cdot Q^{(1),(\mathbb{T})}\mathbf{a}^1, \tag{9.4}$$

then (9.3) follows from (8.14), (8.15) and (8.30).

To achieve the decomposition (9.1), as in (8.36) in Section 8.2, we start with a decomposition on $G^{(i)}_{k\,l}$.

Definition 9.1 *As in Lemma 4.3, we use the notation* $(i\,\mathbb{T})$ *for* $(\{i\}\cup\mathbb{T})$*. For* $1\le i\le N$*,* $i\notin\mathbb{S}=\{i_1,i_2,\dots,i_s\}$ *and* $\mathbb{T}\subset\mathbb{S}$*, we define*

$$\mathcal{G}^{(i),\mathbb{S},(\mathbb{T})}_{k\,l}\equiv\sum_{\mathbb{U}:\mathbb{T}\subset\mathbb{U}\subset\mathbb{S}\backslash\{k,\ l\}}(-1)^{|\mathbb{U}|-|\mathbb{T}|}G^{(i\,\mathbb{T})}_{k\,l}. \tag{9.5}$$

For example, by (8.36), for the case $\mathbb{S}=\{2,3\}$ and $i=1$, we have $\mathcal{G}^{(1),\{2,3\},(\mathbb{T})}_{k\,l}=Q^{(1),(\mathbb{T})}_{k\,l}$; for the case $\mathbb{S}=\{2\}$ and $i=1$, from (8.11) and (8.12) we have $\mathcal{G}^{(1),\{2\},(\mathbb{T})}_{k\,l}=P^{(1),(\mathbb{T})}_{k\,l}$.

From this definition one can easily check that

1. $$\mathcal{G}^{(i),\mathbb{S},(\mathbb{T})}_{k\,l}=0,\quad \text{if } k \text{ or } l\in\mathbb{T}\cup\{i\}. \tag{9.6}$$

2. For $k,l\notin\mathbb{T}\cup\{i\}$,
$$\mathcal{G}^{(i),\mathbb{S},(\mathbb{T})}_{k\,l}=\mathcal{G}^{(i),\mathbb{S}\backslash\{k,\ l\},(\mathbb{T})}_{k\,l}. \tag{9.7}$$

3. $\mathcal{G}^{(i),\mathbb{S},(\mathbb{T})}_{k\,l}$ is independent of the rows or columns of H in $\{i\}\cup\mathbb{T}$, i.e.,
$$\frac{\partial \mathcal{G}^{(i),\mathbb{S},(\mathbb{T})}_{k\,l}}{\partial \mathbf{a}^a_b}=0,\ \ \frac{\partial \mathcal{G}^{(i),\mathbb{S},(\mathbb{T})}_{k\,l}}{\partial \overline{\mathbf{a}^a_b}}=0,\ \ a\in\{i\}\cup\mathbb{T},\ \ 1\le b\le N. \tag{9.8}$$

4. All quantities defined so far depend on the initial matrix H, omitted in our notations. If we wish to specify which matrix is being considered, we will insert the matrix. For example, $\mathcal{G}^{(i),\mathbb{S}\backslash\mathbb{T},\emptyset}_{k\,l}(H^{(\mathbb{T})})$ means it is defined w.r.t. $H^{(\mathbb{T})}$ which is the $N-|\mathbb{T}|$ by $N-|\mathbb{T}|$ minor of H after removing the rows and columns in $\mathbb{T}$. Clearly, we have the relation
$$\mathcal{G}^{(i),\mathbb{S},(\mathbb{T})}_{k\,l}(H)=\mathcal{G}^{(i),\mathbb{S}\backslash\mathbb{T},\emptyset}_{k\,l}(H^{(\mathbb{T})}). \tag{9.9}$$

With these definitions, we can decompose $G^{(i)}_{k\,l}$ as follows.

Lemma 9.2 *For fixed i, $\mathbb{S} = \{i_1, i_2, \ldots, i_s\}$ such that $i \notin \mathbb{S}$, we have the decomposition*

$$G^{(i)}_{k\,l} = \sum_{\mathbb{T}\subset\mathbb{S}} \mathcal{G}^{(i),\mathbb{S},(\mathbb{T})}_{k\,l}. \tag{9.10}$$

Proof. Using the definition (9.5), we have

$$\sum_{\mathbb{T}\subset\mathbb{S}} \mathcal{G}^{(i),\mathbb{S},(\mathbb{T})}_{k\,l} = \sum_{\mathbb{T}\subset\mathbb{S}} \left(\sum_{\mathbb{U}:\mathbb{T}\subset\mathbb{U}\subset\mathbb{S}\backslash\{k,\,l\}} (-1)^{|\mathbb{U}|-|\mathbb{T}|} G^{(1\,\mathbb{U})}_{k\,l} \right) = \sum_{\mathbb{U}\subset\mathbb{S}\backslash\{k,\,l\}} \left(\sum_{\mathbb{T}\subset\mathbb{U}} (-1)^{|\mathbb{U}|-|\mathbb{T}|} \right) G^{(1\,\mathbb{U})}_{k\,l}. \tag{9.11}$$

Since $\sum_{\mathbb{T}\subset\mathbb{U}}(-1)^{|\mathbb{U}|-|\mathbb{T}|} = 0$ unless $\mathbb{U} = \emptyset$, we obtain (9.10) and this concludes Lemma 9.2. □

For the special case $i = 1$ and $\mathbb{S} = \{2, 3\}$, $\mathcal{G}^{(i),\mathbb{S},(\mathbb{T})}_{k\,l} = Q^{(i),(\mathbb{T})}_{kl}$ satisfies the estimate (8.45). We now prove a general form of this estimate on $\mathcal{G}^{(i),\mathbb{S},(\mathbb{T})}_{k\,l}$.

Lemma 9.3 *Let $1 \le i \le N$ and $\mathbb{T} \subset \mathbb{S} = \{i_1, i_2, \ldots, i_s\}$ such that $i \notin \mathbb{S}$. Then there exists a constant C, depending only on s, such that*

$$\left|\mathcal{G}^{(i),\mathbb{S},(\mathbb{T})}_{k\,l}\right| \le C\left(\mathbf{1}(\mathbb{T}\cup\{k\} = \mathbb{S}, k = l) + X^{|\mathbb{S}\backslash(\mathbb{T}\cup\{k,\,l\})|+1}\right), \quad \text{in } \Omega^c, \tag{9.12}$$

for sufficiently large N depending only on s.

This lemma is the basic estimate for a power counting argument. It shows that the off-diagonal elements of $\mathcal{G}^{(i),\mathbb{S},(\mathbb{T})}_{k\,l}$ are small by a certain power of X, which is our small parameter, depending on the size of the sets $\mathbb{S}$ and $\mathbb{T}$. The diagonal elements, when not zero by definition, are estimated by 1 (first term in (9.12)), but their contribution to the moments of $\mathcal{Z}^{(i),\mathbb{S},(\mathbb{T})}$ will be small since $k = l$ reduces the double sum in (9.1) to a single sum.

Proof of Lemma 9.3. For $k = l$, the estimate (9.12) follows directly from (9.5) and (8.5). We can thus assume that $k \neq l$ throughout the proof of this lemma. The argument consists of two parts. First we prove a representation formula (Lemma 9.4) that asserts that $\mathcal{G}^{(i),\mathbb{S},(\mathbb{T})}_{k\,l}$ is a certain rational function involving resolvent matrix elements of H and some of its minors. The denominators in this rational function are products of diagonal elements of resolvents and the numerators are products of off-diagonal matrix elements. In the second step we will estimate these rational functions, using that the diagonal elements of the resolvent are typically separated away from zero and the off-diagonal elements are small by a factor X.

For the precise argument, we start with the cases:

$$\mathbb{T} = \emptyset, \;\; k, l \notin \mathbb{S} \text{ and } \mathbb{S} \neq \emptyset. \tag{9.13}$$

The special case $\mathbb{S} = \{2\}$ can be proved by the representation (8.11) and Lemma 8.1. The case $\mathbb{S} = \{2, 3\}$ was proved in (8.45). These examples show that $\mathcal{G}^{(i),\mathbb{S},(\mathbb{T})}_{k\,l}$ can be written as the finite sum of the terms of the form:

$$\frac{G_o G_o \cdots G_o}{G_d G_d \cdots G_d}, \tag{9.14}$$

where G_o are off-diagonal elements of some $G^{(\mathbb{U})}$ and G_d are diagonal elements. Furthermore, in each term, the number of the off-diagonal elements in the numerator is strictly greater than $s = |\mathbb{S}|$ but less than 4^s. The number of the diagonal elements in the denominator is also less than $4^{|\mathbb{S}|}$.

The Green function $G^{(i,\mathbb{T})}_{k\,l}$ can be viewed as a function from the vector space of matrices. This motivates the following definition.

Definition 9.2 *Denote by $\mathcal{X}_K$ the space of $K \times K$ matrices and $\mathcal{X} = \cup_{K=1}^{\infty} \mathcal{X}_K$. Define $\mathcal{Y}$ as the set of functions from $\mathcal{X}$ to the complex numbers. For any $\mathbb{S} = \{i_1, i_2, \ldots, i_s\}$ and for any $i, k, l \notin \mathbb{S}$, define the set of off-diagonal matrix elements considered as functions of matrices:*

$$\mathcal{A}^{(i),\mathbb{S}}_{k\,l} \equiv \Big\{ f \in \mathcal{Y} : f(W) = G^{(i\,\mathbb{U})}_{jj'}(W), \text{ for some } j \neq j', \ j, j' \in \mathbb{S} \cup \{k, l\}, \ \mathbb{U} \subset \mathbb{S} \Big\}, \tag{9.15}$$

where $W \in \mathcal{X}_K$ for some K. Similarly, we define the set of diagonal matrix elements:

$$\mathcal{B}^{(i),\mathbb{S}}_{k\,l} \equiv \Big\{ f \in \mathcal{Y} : f(W) = G^{(i\,\mathbb{U})}_{jj}(W) : j \in \mathbb{S} \cup \{k, l\}, \quad \mathbb{U} \subset \mathbb{S} \Big\}. \tag{9.16}$$

Furthermore we define $\mathcal{C}^{(i),\mathbb{S}}_{k\,l}$ for all k, l as

$$\mathcal{C}^{(i),\mathbb{S}}_{k\,l} \equiv \Big\{ F \in \mathcal{Y} \text{ is a finite sum of functions of the form } \pm \frac{f_1 f_2 \cdots f_m}{g_1 g_2 \cdots g_{m'}} : \\ f_\alpha \in \mathcal{A}^{(i),\mathbb{S}}_{k\,l}, 1 \leq \alpha \leq m; \ g_\beta \in \mathcal{B}^{(i),\mathbb{S}}_{k\,l}, \ 1 \leq \beta \leq m'; \ s+1 \leq m \leq 4^s, \ 0 \leq m' \leq 4^s \Big\}, \tag{9.17}$$

where $s = |\mathbb{S}|$.

Notice the important condition $m \geq |\mathbb{S}| + 1$ in the definition of $\mathcal{C}^{(i),\mathbb{S}}_{k\,l}$. Since off-diagonal matrix elements are typically small, this requirement will guarantee the smallness of $\mathcal{C}^{(i),\mathbb{S}}_{k\,l}$ as a certain power of X.

With these notations, the equation (8.41) asserts that for $k, l \notin \{2, 3\}$, there is a function $F^{(1),\{2,3\}}_{k,l} \in \mathcal{C}^{(1),\{2,3\}}_{k\,l}$ such that

$$\mathcal{G}^{(1),\{2,3\},\emptyset}_{k\,l} = F^{(1),\{2,3\}}_{k,l}. \tag{9.18}$$

The general case is the following lemma.

Lemma 9.4 *For any $\mathbb{S} = \{i_1, i_2, \ldots, i_s\}$ with $s > 0$ and $i, k, l \notin \mathbb{S}$, there exists a function $F^{(i),\mathbb{S}}_{k,l} \in \mathcal{C}^{(i),\mathbb{S}}_{k\,l}$ such that*

$$\mathcal{G}^{(i),\mathbb{S},\emptyset}_{k\,l} = F^{(i),\mathbb{S}}_{k,l}. \tag{9.19}$$

Proof of Lemma 9.4: By symmetry, we only need to prove the cases that

$$i = 1, \quad \mathbb{S} = \{2, 3, \ldots, s+1\}.$$

To prove this case, we argue by induction on s. For $s = 1$ or 2, Lemma 9.4 was proved in (8.11) and (8.41) (cf. (9.18)). Suppose that Lemma 9.4 is correct for $s = n - 1 \geq 1$ and $F^{(1),\{2,\ldots,n\}}_{k,l} \in \mathcal{C}^{(1),\{2,\ldots,n\}}_{k,l}$ is the function satisfying (9.19) for $i = 1$ and $\mathbb{S} = \{2, 3, \ldots, n\}$

Now let $i = 1$, $\mathbb{S} = \{2, \dots, n+1\}$ and $k, l \notin \{1, \dots, n+1\}$. By the induction assumption,

$$\mathcal{G}_{k\,l}^{(1),\{2,\dots,n\},\emptyset} = F_{k,l}^{(1),\{2,\dots,n\}} \tag{9.20}$$

with $F_{k,l}^{(1),\{2,\dots,n\}}$ a finite sum of elements of the form

$$\pm \left(\prod_{\alpha=1}^{m} G_{j_\alpha j'_\alpha}^{(1\,\mathbb{U}_\alpha)} \right) \left(\prod_{\beta=1}^{m'} G_{j_{m+\beta} j_{m+\beta}}^{(1\,\mathbb{U}_{m+\beta})} \right)^{-1}, \tag{9.21}$$

where $\mathbb{U}_\alpha \subset \mathbb{S}$, $n \le m \le 4^{n-1}$, $0 \le m' \le 4^{n-1}$ and

$$j_\alpha, j\,'_\alpha, j_{m+\beta} \in \{2, 3, \dots, n\} \cup \{k\} \cup \{l\}.$$

By definition of $\mathcal{G}_{k\,l}^{(1),\mathbb{S},(\mathbb{T})}$ in (9.5), we have

$$\mathcal{G}_{kl}^{(1),\{2,\dots.,n+1\},\emptyset} = \mathcal{G}_{kl}^{(1),\{2,..n\},\emptyset} - \mathcal{G}_{kl}^{(1),\{2,\dots,n,n+1\},(n+1)} \tag{9.22}$$

Combining (9.9) with (9.20), we have

$$\mathcal{G}_{k\,l}^{(1),\{2,3\dots,n,n+1\},(n+1)}(W) = F_{k,l}^{(1),\{2,\dots,n\}}(W^{(n+1)}), \tag{9.23}$$

where $W^{(n+1)}$ is the minor of W with the $(n+1)$-th row and $(n+1)$-th column removed.

We can remove the dependence on the $(n+1)$-th row by the procedure in (8.37)-(8.39). Using (4.15), (4.16) and the notation:

$$(1\,\mathbb{U}\,n+1) = (\{1, n+1\} \cup \mathbb{U}), \tag{9.24}$$

we have the expansion

$$G_{j_\alpha\, j'_\alpha}^{(1\,\mathbb{U}_\alpha)} = G_{j_\alpha\, j'_\alpha}^{(1\,\mathbb{U}_\alpha n+1)} + \frac{G_{j_\alpha\, n+1}^{(1\,\mathbb{U}_\alpha)} G_{n+1\, j'_\alpha}^{(1\,\mathbb{U}_\alpha)}}{G_{n+1\, n+1}^{(1\,\mathbb{U}_\alpha)}}, \qquad 1 \le \alpha \le m \tag{9.25}$$

and

$$\frac{1}{G_{j_\beta\, j_\beta}^{(1\,\mathbb{U}_\beta)}} = \frac{1}{G_{j_\beta\, j_\beta}^{(1\,\mathbb{U}_\beta\, n+1)}} - \frac{G_{j_\beta\, n+1}^{(i\,\mathbb{U}_\beta)} G_{n+1\, j_\beta}^{(i\,\mathbb{U}_\beta)}}{G_{j_\beta\, j_\beta}^{(i\,\mathbb{U}_\beta)} G_{j_\beta\, j_\beta}^{(i\,\mathbb{U}_\beta\, n+1)} G_{n+1\, n+1}^{(i\,\mathbb{U}_\beta)}}, \quad m+1 \le \beta \le m+k'. \tag{9.26}$$

We note that the first term on the r.h.s of (9.25) is exactly the Green function on the l.h.s of (9.25) except that there is an additional superscript $n+1$; the similar comment applies to (9.26).

Inserting (9.25) and (9.26) into (9.21) and expanding it, we obtain that (9.21) is equal to

$$\pm \left(\prod_{\alpha=1}^{m} G_{j_\alpha j'_\alpha}^{(1\,\mathbb{U}_\alpha\, n+1)} \right) \left(\prod_{\beta=1}^{m'} G_{j_{m+\beta}\, j_{m+\beta}}^{(1\,\mathbb{U}_{m+\beta}\, n+1)} \right)^{-1} + \text{ other terms.} \tag{9.27}$$

Here the first term in (9.27) is the product of the first terms on the right side of (9.25) and (9.26) and it is the same as (9.21) except that there is an additional superscript $n+1$. One can see that the other terms in (9.27) are elements in $\mathcal{C}_{k\,l}^{(1),\{2,\dots,n+1\}}$, i.e., the number of the off diagonal terms in the numerator is now at least $n+1$. Since this procedure can be applied to each term in $F_{k,l}^{(1),\{2,\ldots,n\}}$, we have proved that there exists an $F \in \mathcal{C}_{k\,l}^{(1),\{2,\ldots,n+1\}}$ such that

$$\begin{aligned}\mathcal{G}_{k\,l}^{(1),\{2,\ldots,n\},\emptyset}(W) = F_{k,l}^{(1),\{2,\ldots,n\}}(W) &= F_{k,l}^{(1),\{2,\ldots,n\}}\left(W^{(n+1)}\right) + F(W) \\ &= \mathcal{G}_{k\,l}^{(1),\{2,\ldots,n+1\},(n+1)}(W) + F(W). \end{aligned} \tag{9.28}$$

By (9.22) and (9.23), we can set $F_{k,l}^{(i),\{2,3\ldots,n,n+1\}}(W) = F(W)$ which is in $\mathcal{C}_{k\,l}^{(1),\{2,3,\ldots,n,n+1\}}$ and we have thus proved Lemma 9.4 by induction. □

Now we start proving the estimates in Lemma 9.3. Using (9.6) and (9.7), we only have to prove (9.12) for the case $k, l \notin \mathbb{S} \cup \{i\}$.

Case 1, $\mathbb{T} = \mathbb{S}$*:* By definition,

$$\mathcal{G}_{k\,l}^{(i),\,\mathbb{S},\,(\mathbb{T})} = G_{k\,l}^{(i\,\mathbb{S})}. \tag{9.29}$$

Then (9.12) in this special case follows from Lemma 8.1.

Case 2, $\mathbb{T} = \emptyset, \quad k, l \notin \mathbb{S}$ *and* $\mathbb{S} \neq \emptyset$*:* By Lemma 9.4 and 8.1, for any $\mathbb{S} \neq \emptyset$ such that $i, k, l \notin \mathbb{S}$, we have

$$|\mathcal{G}_{k\,l}^{(i),\mathbb{S},\emptyset}| \leq C \frac{\left(\max_{\mathbb{U}\subset\mathbb{S}, j\neq j'} |G_{jj'}^{(i\,\mathbb{U})}|\right)^{s+1}}{\left(\min_{\mathbb{U}\subset\mathbb{S}, j} |G_{jj}^{(i\,\mathbb{U})}|\right)^{4^s}} \leq C X^{s+1}, \tag{9.30}$$

where C depends on $s = |\mathbb{S}|$.

Case 3, $\mathbb{T} \neq \emptyset, \quad \mathbb{T} \subset \mathbb{S}, \quad \mathbb{T} \neq \mathbb{S}, \quad k, l \notin \mathbb{S}$ *and* $\mathbb{S} \neq \emptyset$*:* By (9.9) and Lemma 9.4, there exists a function $F_{k,l}^{(i),\mathbb{S}\setminus\mathbb{T}} \in \mathcal{C}_{k\,l}^{(i),\mathbb{S}\setminus\mathbb{T}}$ (see (9.17)) such that

$$\mathcal{G}_{k\,l}^{(i),\mathbb{S},(\mathbb{T})}(H) = \mathcal{G}_{k\,l}^{(i),\mathbb{S}\setminus\mathbb{T}}(H^{(\mathbb{T})}) = F_{k,l}^{(i),\mathbb{S}\setminus\mathbb{T}}(H^{(\mathbb{T})}), \tag{9.31}$$

where $H^{(\mathbb{T})}$ is the $N - |\mathbb{T}|$ by $N - |\mathbb{T}|$ minor of H after removing the rows and columns in $\mathbb{T}$. Thus $\mathcal{G}_{k\,l}^{(i),\mathbb{S},(\mathbb{T})}$ is given by the function $F_{k,l}^{(i),\mathbb{S}\setminus\mathbb{T}}$ with all Green functions $G_{jj'}^{(\mathbb{U})}$ in the definition of $F_{k,l}^{(i),\mathbb{S}\setminus\mathbb{T}}$ replaced by $G_{jj'}^{(\mathbb{U}\cup\mathbb{T})}$. From (9.30) we have

$$|\mathcal{G}_{k\,l}^{(i),\mathbb{S},(\mathbb{T})}| \leq C \frac{\left(\max_{\mathbb{U}\subset\mathbb{S}\setminus\mathbb{T}, j\neq j'} |G_{jj'}^{(i\,\mathbb{U}\cup\mathbb{T})}|\right)^{|\mathbb{S}\setminus\mathbb{T}|+1}}{\left(\min_{\mathbb{U}\subset\mathbb{S}\setminus\mathbb{T}, j} |G_{jj}^{(i\,\mathbb{U}\cup\mathbb{T})}|\right)^{(4^{|\mathbb{S}\setminus\mathbb{T}|})}}. \tag{9.32}$$

Using Lemma 8.1, we have that

$$|\mathcal{G}_{k\,l}^{(i),\mathbb{S},(\mathbb{T})}| \leq C X^{|\mathbb{S}\setminus\mathbb{T}|+1}, \quad \text{in } \Omega^c \tag{9.33}$$

where C depends on s. We have thus proved (9.12) for the Case 3 and this completes the proof of Lemma 9.3. □

Proof of Lemma 9.1. The decomposition (9.1) follows from (9.10) and (9.2) is a direct consequence of (9.8). The estimate (9.3) can be proved in the same way as in the proof of Lemma 8.2 using the following three ingredients: (1) The bounds on $\left|\mathcal{G}^{(i),\mathbb{S},(\mathbb{T})}_{k\,l}\right|$ in (9.12). (2) The large deviation estimate in Lemma 4.4. (3) The trivial bound $|\mathcal{G}^{(i),\mathbb{S},(\mathbb{T})}_{k\,l}| \leq C/\eta \leq CN$ where C depends on $|\mathbb{S}|$. This concludes the proof of Lemma 9.1. □

Proof of Lemma 5.2. We first introduce the following notations which will be useful for the expansion of the p-th moment of $\left|\sum_{i=1}^N Z_i\right|$ in (5.5).

Definition 9.3 *1. Let $\mathbf{V} = \langle v_1, v_2, \ldots, v_p\rangle$ be a p dimensional vector such that $v_i = 0$ or 1 for $1 \leq i \leq p$.*

2. Let $\mathbf{S} = \langle \alpha_1, \alpha_2, \ldots, \alpha_p\rangle$ be a p dimensional vector such that $1 \leq \alpha_i \leq N$ for $1 \leq i \leq p$.

3. Denote by $\mathbb{S}$ the set consisting of elements α_j which is a component of $\mathbf{S}$.
We define

$$A(\mathbf{S},\mathbf{V}) = \mathbb{E}\prod_{j=1}^{p}\left(\mathbf{B}^{v_j} Z_{\alpha_j}\right), \quad \mathbf{B}^1(a+ib) = a - ib, \quad \mathbf{B}^0(a+ib) = a+ib, \tag{9.34}$$

where $\mathbf{B}$ is the complex conjugate operator.

Through the rest of this section, $\mathbb{S}$ is always the set generated by $\mathbf{S}$. Notice that $|\mathbb{S}| = s \leq p$ where p is the number of components in $\mathbf{S}$. With these notations, we can estimate $\mathbb{E}\,|\sum_i Z_i|^p$ by

$$\mathbb{E}\left|\sum_{i=1}^N Z_i\right|^p \leq \sum_{\mathbf{S}}\sum_{\mathbf{V}}|A(\mathbf{S},\mathbf{V})| \leq C_p \sum_{s\leq p} N^s \max_{\mathbf{S},\mathbf{V}:|\mathbb{S}|=s}|A(\mathbf{S},\mathbf{V})|\,, \tag{9.35}$$

where we sum up $\mathbf{S}$ and $\mathbf{V}$ under the conditions in Definition 9.3. Lemma 5.2 is now a simple consequence of the following estimate on $|A(\mathbf{S},\mathbf{V})|$.

Lemma 9.5 *Let $\mathbb{S}$, $\mathbf{S}$ and $\mathbf{V}$ satisfy the conditions in Definition 9.3. With $A(\mathbf{S},\mathbf{V})$ defined in (9.34), there exists a constant $C>0$ depending on p such that*

$$|A(\mathbf{S},\mathbf{V})| \leq C\left((\log N)^{3+2\alpha}\right)^p N^{p-s}X^{2p}, \tag{9.36}$$

for sufficiently large N depending only on p.

Proof. Let $\mathbb{S}_i, 1 \leq i \leq p$, denote the set $\mathbb{S}_i = \mathbb{S}\backslash\{\alpha_i\}$. Using (9.1), we expand $A(\mathbf{S},\mathbf{V})$ as

$$\begin{aligned} A(\mathbf{S},\mathbf{V}) &= \mathbb{E}\sum_{\mathbb{T}_1\subset\mathbb{S}_1}\ldots\sum_{\mathbb{T}_p\subset\mathbb{S}_p} A(\mathbb{T}_1,\mathbb{T}_2,\ldots\mathbb{T}_p,\mathbf{V}), \qquad (9.37)\\ A(\mathbb{T}_1,\mathbb{T}_2,\ldots\mathbb{T}_p,\mathbf{V}) &\equiv \left(\mathbf{B}^{v_1}\mathbb{IE}_{\alpha_1}\mathcal{Z}^{(\alpha_1),\mathbb{S}_1,(\mathbb{T}_1)}\right)\left(\mathbf{B}^{v_2}\mathbb{IE}_{\alpha_2}\mathcal{Z}^{(\alpha_2),\mathbb{S}_2,(\mathbb{T}_2)}\right)\cdots \end{aligned}$$

From the Schwarz inequality, (9.3) and $|\mathbb{S}_i| = s-1$, we obtain that

$$|\mathbb{E}A(\mathbb{T}_1,\mathbb{T}_2,\ldots\mathbb{T}_p,\mathbf{V})| \leq C\left((\log N)^{3+2\alpha}\right)^p X^{\left(ps-\sum_{i=1}^p|\mathbb{T}_i|\right)}, \tag{9.38}$$

where C depends on p. Suppose that

$$\sum_{i=1}^p|\mathbb{T}_i| \leq sp-2s. \tag{9.39}$$

Using (8.20), i.e., $X^2 \geq (\log N)^2/M \geq 1/N$, we have

$$|\mathbb{E}A(\mathbb{T}_1, \mathbb{T}_2, \dots \mathbb{T}_p, \mathbf{V})| \leq C\left((\log N)^{3+2\alpha}\right)^p X^{2s} \leq C\left((\log N)^{3+2\alpha}\right)^p N^{p-s} X^{2p}. \tag{9.40}$$

It remains to estimate $\mathbb{E}A(\mathbb{T}_1, \mathbb{T}_2, \dots \mathbb{T}_p, \mathbf{V})$ for the cases that

$$\sum_{i=1}^{p} |\mathbb{T}_i| \geq s\,p - 2s + 1. \tag{9.41}$$

For $\gamma \in \mathbb{S}$, denote n_γ to be the number of times that γ appears in $\{\alpha_1\} \cup \mathbb{T}_1$, $\{\alpha_2\} \cup \mathbb{T}_2, \dots$ and $\{\alpha_p\} \cup \mathbb{T}_p$, i.e.,

$$n_\gamma = \sum_{k=1}^{p} \mathbf{1}(\gamma \in \{\alpha_k\} \cup \mathbb{T}_k).$$

By definition, $n_\gamma \geq 1$. Similarly, we define m_γ to be the number of times that γ appears in $\langle \alpha_1, \alpha_2, \dots \alpha_p \rangle$, i.e.,

$$m_\gamma = \sum_{k=1}^{p} \mathbf{1}(\gamma = \alpha_k).$$

Let $x = |\{\gamma \in \mathbb{S} : n_\gamma = p\}|$ and $y = |\{\gamma \in \mathbb{S} : m_\gamma = 1\}|$. Since for each fixed i, $\alpha_i \notin \mathbb{T}_i$, then with (9.41) and the definition of n_γ,

$$(p-1)(s-x) + xp \geq \sum_{\gamma \in \mathbb{S}} n_\gamma = \sum_{i=1}^{p} |\{\alpha_i\} \cup \mathbb{T}_i| = p + \sum_{i=1}^{p} |\mathbb{T}_i| \geq sp - 2s + p + 1. \tag{9.42}$$

By definition of m_γ, we have

$$y + 2(s-y) \leq \sum_{\gamma \in \mathbb{S}} m_\gamma = p. \tag{9.43}$$

From the last two inequalities, we have $x + y \geq s + 1$ and thus there exists a $\gamma \in \mathbb{S}$ such that

$$n_\gamma = p \ \text{ and } \ m_\gamma = 1. \tag{9.44}$$

Without loss of generality, we assume that $\gamma = \alpha_1$. Then using (9.44), we know

$$\gamma \neq \alpha_k, \ \ \gamma \in \mathbb{T}_k, \quad \text{ if } \ k \neq 1. \tag{9.45}$$

Then with (9.45), the decomposition $\mathcal{Z}^{(i),\mathbb{S},(\mathbb{T})} \equiv \sum_{k,l} \overline{\mathbf{a}}^i_k \mathcal{G}^{(i),\mathbb{S},(\mathbb{T})}_{k\,l} \mathbf{a}^i_l$ (9.1) and the property that $\mathcal{G}^{(i),\mathbb{S},(\mathbb{T})}_{k\,l}$ is independent of the row or columns of H in $\{i\} \cup \mathbb{T}$ (9.2), we have that for $k \neq 1$, the $\mathcal{Z}^{\alpha_k,\mathbb{S}_k,(\mathbb{T}_k)}$ is independent of $\mathbf{a}^\gamma$. By the definition of $\mathbb{IE}$, for $k = 1$, we also have

$$\mathbb{E}_{\mathbf{a}^\gamma} \mathbb{IE}_{\mathbf{a}^{\alpha_1}} \mathcal{Z}^{(\alpha_1),\mathbb{S}_1,(\mathbb{T}_1)} = \mathbb{E}_{\mathbf{a}^\gamma} \mathbb{IE}_{\mathbf{a}^\gamma} \mathcal{Z}^{(\gamma),\mathbb{S}_1,(\mathbb{T}_1)} = 0. \tag{9.46}$$

Therefore, under the assumption (9.41) we have

$$\begin{aligned} \mathbb{E}A(\mathbb{T}_1, \mathbb{T}_2, \dots \mathbb{T}_s, \mathbf{V}) &= \mathbb{E}\left(\mathbf{B}^{v_1} \mathbb{IE}_{\mathbf{a}^{\alpha_1}} \mathcal{Z}^{(\alpha_1),\mathbb{S}_1,(\mathbb{T}_1)}\right)\left(\mathbf{B}^{v_2} \mathbb{IE}_{\mathbf{a}^{\alpha_2}} \mathcal{Z}^{(\alpha_2),\mathbb{S}_2,(\mathbb{T}_2)}\right) \cdots \\ &= \mathbb{E}\ \left(\mathbb{E}_{\mathbf{a}^\gamma} \mathbf{B}^{v_1} \mathbb{IE}_{\mathbf{a}^{\alpha_1}} \mathcal{Z}^{(\alpha_1),\mathbb{S}_1,(\mathbb{T}_1)}\right)\left(\mathbf{B}^{v_2} \mathbb{IE}_{\mathbf{a}^{\alpha_2}} \mathcal{Z}^{(\alpha_2),\mathbb{S}_2,(\mathbb{T}_2)}\right) \cdots = 0. \end{aligned}$$

Combining this identity with (9.40), we obtain (9.36) and thus conclude Lemma 9.5. □

Acknowledgement. The authors thank Terry Tao for his helpful comments on an earlier version of this manuscript.

References

[1] Anderson, G., Guionnet, A., Zeitouni, O.: *An Introduction to Random Matrices.* Studies in Advanced Mathematics, **118**, Cambridge University Press, 2009.

[2] Anderson, G.; Zeitouni, O. : A CLT for a band matrix model. *Probab. Theory Related Fields* **134** (2006), no. 2, 283–338.

[3] Ben Arous, G., Péché, S.: Universality of local eigenvalue statistics for some sample covariance matrices. *Comm. Pure Appl. Math.* **LVIII.** (2005), 1–42.

[4] Bleher, P., Its, A.: Semiclassical asymptotics of orthogonal polynomials, Riemann-Hilbert problem, and universality in the matrix model. *Ann. of Math.* **150** (1999): 185–266.

[5] Deift, P.: Orthogonal polynomials and random matrices: a Riemann-Hilbert approach. *Courant Lecture Notes in Mathematics* **3**, American Mathematical Society, Providence, RI, 1999.

[6] Deift, P., Gioev, D.: Random Matrix Theory: Invariant Ensembles and Universality. *Courant Lecture Notes in Mathematics* **18**, American Mathematical Society, Providence, RI, 2009.

[7] Deift, P., Kriecherbauer, T., McLaughlin, K.T-R, Venakides, S., Zhou, X.: Uniform asymptotics for polynomials orthogonal with respect to varying exponential weights and applications to universality questions in random matrix theory. *Comm. Pure Appl. Math.* **52** (1999):1335–1425.

[8] Deift, P., Kriecherbauer, T., McLaughlin, K.T-R, Venakides, S., Zhou, X.: Strong asymptotics of orthogonal polynomials with respect to exponential weights. *Comm. Pure Appl. Math.* **52** (1999): 1491–1552.

[9] Disertori, M., Pinson, H., Spencer, T.: Density of states for random band matrices. *Commun. Math. Phys.* **232**, 83–124 (2002)

[10] Dyson, F.J.: A Brownian-motion model for the eigenvalues of a random matrix. *J. Math. Phys.* **3**, 1191–1198 (1962).

[11] Erdős, L., Schlein, B., Yau, H.-T.: Semicircle law on short scales and delocalization of eigenvectors for Wigner random matrices. *Ann. Probab.* **37**, No. 3, 815–852 (2008).

[12] Erdős, L., Schlein, B., Yau, H.-T.: Local semicircle law and complete delocalization for Wigner random matrices. *Commun. Math. Phys.* **287**, 641–655 (2009).

[13] Erdős, L., Schlein, B., Yau, H.-T.: Wegner estimate and level repulsion for Wigner random matrices. *Int. Math. Res. Notices.* **2010**, No. 3, 436-479 (2010).

[14] Erdős, L., Schlein, B., Yau, H.-T.: Universality of random matrices and local relaxation flow. To appear in *Inv. Math.* Preprint arXiv:0907.5605

[15] Erdős, L., Ramirez, J., Schlein, B., Yau, H.-T.: *Universality of sine-kernel for Wigner matrices with a small Gaussian perturbation.* Electr. J. Prob. **15**, Paper 18, 526–604 (2010).

[16] Erdős, L., Péché, G., Ramírez, J., Schlein, B., and Yau, H.-T., Bulk universality for Wigner matrices. *Comm. Pure Appl. Math.* **63**, No. 7, 895-925 (2010).

[17] Erdős, L., Ramírez, J., Schlein, B., Tao, T., Vu, V. and Yau, H.-T., Bulk universality for Wigner hermitian matrices with subexponential decay. *Math. Res. Lett.* **17** (2010), no. 4, 667–674.

[18] Erdős, L., Schlein, B., Yau, H.-T., Yin, J.: The local relaxation flow approach to universality of the local statistics for random matrices. To appear in *Annales Inst. H. Poincaré (B), Probability and Statistics.* Preprint arXiv:0911.3687

[19] Erdős, L., Yau, H.-T., Yin, J.: Bulk universality for generalized Wigner matrices. Preprint arXiv:1001.3453

[20] Forrester, P. J.: Log-gases and random matrices. London Mathematical Society Monographs, 2010.

[21] Guionnet, A.: Large deviation upper bounds and central limit theorems for band matrices, *Ann. Inst. H. Poincaré Probab. Statist* **38** , (2002), pp. 341-384.

[22] Hanson, D.L., Wright, F.T.: A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Math. Stat.* **42** (1971), no.3, 1079-1083.

[23] Johansson, K.: Universality of the local spacing distribution in certain ensembles of Hermitian Wigner matrices. *Comm. Math. Phys.* **215** (2001), no.3. 683–705.

[24] Mehta, M.L.: Random Matrices. Academic Press, New York, 1991.

[25] Pastur, L., Shcherbina M.: Bulk universality and related properties of Hermitian matrix models. *J. Stat. Phys.* **130** (2008), no.2., 205-250.

[26] Péché, S: Universality in the bulk of the spectrum for complex sample covariance matrices, Preprint, arXiv:0912.2493.

[27] Spencer, T.: Review article on random band matrices. Draft in preparation.

[28] Tao, T. and Vu, V.: Random matrices: Universality of the local eigenvalue statistics, *Acta Math.*, **206** (2011), Number 1, 127-204 Preprint arXiv:0906.0510.

[29] Tao, T. and Vu, V.: Random covariance matrices: Universality of local statistics of eigenvalues. Preprint. arXiv:0912.0966

[30] Tao, T. and Vu, V.: Random matrices: Universality of local eigenvalue statistics up to the edge. *Comm. Math. Phys.* **298** (2010), no. 2, 549???572.

[31] Tao, T. and Vu, V.: Random matrices: Localization of the eigenvalues and the necessity of four moments. Preprint. arXiv:1005.2901

[32] Wigner, E.: Characteristic vectors of bordered matrices with infinite dimensions. *Ann. of Math.* **62** (1955), 548-564.